

8-04-04

IFW

PTO/SB/21 (04-04)

	Application Number	10/626,049
	Filing Date	July 23, 2003
	First Named Inventor	Hosoya, Mutsumi
	Art Unit	2186
	Examiner Name	Unassigned
	Attorney Docket Number	16869P-079400US
Total Number of Pages in This Submission		13

ENCLOSURES (Check all that apply)		
<input checked="" type="checkbox"/> Fee Transmittal Form <input type="checkbox"/> Fee Attached <input type="checkbox"/> Amendment/Reply <input type="checkbox"/> After Final <input type="checkbox"/> Affidavits/declaration(s) <input type="checkbox"/> Extension of Time Request <input type="checkbox"/> Express Abandonment Request <input type="checkbox"/> Information Disclosure Statement <input type="checkbox"/> Certified Copy of Priority Document(s) <input type="checkbox"/> Response to Missing Parts/Incomplete Application <input type="checkbox"/> Response to Missing Parts under 37 CFR 1.52 or 1.53	<input type="checkbox"/> Drawing(s) <input type="checkbox"/> Licensing-related Papers <input checked="" type="checkbox"/> Petition to Make Special <input type="checkbox"/> Petition to Convert to a Provisional Application <input type="checkbox"/> Power of Attorney, Revocation <input type="checkbox"/> Change of Correspondence Address <input type="checkbox"/> Terminal Disclaimer <input type="checkbox"/> Request for Refund <input type="checkbox"/> CD, Number of CD(s) _____	<input type="checkbox"/> After Allowance Communication to Technology Center (TC) <input type="checkbox"/> Appeal Communication to Board of Appeals and Interferences <input type="checkbox"/> Appeal Communication to TC (Appeal Notice, Brief, Reply Brief) <input type="checkbox"/> Proprietary Information <input type="checkbox"/> Status Letter <input checked="" type="checkbox"/> Other Enclosure(s) (please identify below): Return Postcard 6 cited references
Remarks: The Commissioner is authorized to charge any additional fees to Deposit Account 20-1430.		
SIGNATURE OF APPLICANT, ATTORNEY, OR AGENT		
Firm or Individual name	Townsend and Townsend and Crew LLP	
	Chun-Pok Leung	Reg. No. 41,405
Signature		
Date	August 2, 2004	

CERTIFICATE OF TRANSMISSION/MAILING			
Express Mail Label: EV 503884768 US			
I hereby certify that this correspondence is being deposited with the United States Postal Service with "Express Mail Post Office to Address" service under 37 CFR 1.10 on this date August 2, 2004 and is addressed to: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on the date shown below.			
Typed or printed name	Joy Salvador		
Signature		Date	August 2, 2004

FEE TRANSMITTAL for FY 2004

Effective 01/01/2003. Patent fees are subject to annual revision.

Significant claims small entity status. See 37 CFR 1.27

TOTAL AMOUNT OF PAYMENT (\$) 130.00

Complete if Known

Application Number	10/626,049
Filing Date	July 23, 2003
First Named Inventor	Hosoya, Mutsumi
Examiner Name	Unassigned
Art Unit	2186
Attorney Docket No.	16869P-079400US

METHOD OF PAYMENT (check all that apply)

☐ Check ☐ Credit Card ☐ Money Order ☐ Other ☐ None
☒ Deposit Account:Deposit
Account
Number

20-1430

Deposit
Account
Name

Townsend and Townsend and Crew LLP

The Director is authorized to: (check all that apply)

☒ Charge fee(s) indicated below ☒ Credit any overpayments☒ Charge any additional fee(s) or any underpayment of fee(s)☐ Charge fee(s) indicated below, except for the filing fee to the above-identified deposit account.

FEE CALCULATION

1. BASIC FILING FEE

Large Entity		Small Entity		Fee Description	Fee Paid
Fee Code	Fee (\$)	Fee Code	Fee (\$)		
1001	770	2001	385	Utility filing fee	
1002	340	2002	170	Design filing fee	
1003	530	2003	265	Plant filing fee	
1004	770	2004	385	Reissue filing fee	
1005	160	2005	80	Provisional filing fee	

SUBTOTAL (1)

(\$0.00)

2. EXTRA CLAIM FEES FOR UTILITY AND REISSUE

Total Claims		Extra Claims		Fee from below		Fee Paid
Independent Claims		** =		X		
Multiple Dependent				X		

Large Entity		Small Entity		Fee Description
Fee Code	Fee (\$)	Fee Code	Fee (\$)	
1202	18	2202	9	Claims in excess of 20
1201	86	2201	43	Independent claims in excess of 3
1203	290	2203	145	Multiple dependent claim, if not paid
1204	86	2204	43	** Reissue independent claims over original patent
1205	18	2205	9	** Reissue claims in excess of 20 and over original patent

SUBTOTAL (2)

(\$0.00)

**or number previously paid, if greater; For Reissues, see above

FEE CALCULATION (continued)

3. ADDITIONAL FEES

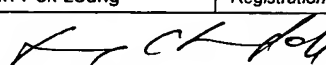
Large Entity		Small Entity		Fee Description	Fee Paid
Fee Code	Fee (\$)	Fee Code	Fee (\$)		
1051	130	2051	65	Surcharge - late filing fee or oath	
1052	50	2052	25	Surcharge - late provisional filing fee or cover sheet	
1053	130	1053	130	Non-English specification	
1812	2,520	1812	2,520	For filing a request for reexamination	
1804	920*	1804	920*	Requesting publication of SIR prior to Examiner action	
1805	1,840*	1805	1,840*	Requesting publication of SIR after Examiner action	
1251	110	2251	55	Extension for reply within first month	
1252	420	2252	210	Extension for reply within second month	
1253	950	2253	475	Extension for reply within third month	
1254	1,480	2254	740	Extension for reply within fourth month	
1255	2,010	2255	1,005	Extension for reply within fifth month	
1401	330	2401	165	Notice of Appeal	
1402	330	2402	165	Filing a brief in support of an appeal	
1403	290	2403	145	Request for oral hearing	
1451	1,510	1451	1,510	Petition to institute a public use proceeding	
1452	110	2452	55	Petition to revive - unavoidable	
1453	1,330	2453	665	Petition to revive - unintentional	
1501	1,330	2501	665	Utility issue fee (or reissue)	
1502	480	2502	240	Design issue fee	
1503	640	2503	320	Plant issue fee	
1460	130	1460	130	Petitions to the Commissioner	130
1807	50	1807	50	Petitions related to provisional applications	
1806	180	1806	180	Submission of Information Disclosure Stmt	
8021	40	8021	40	Recording each patent assignment per property (times number of properties)	
1809	770	2809	385	Filing a submission after final rejection (37 CFR § 1.129(a))	
1810	770	2810	385	For each additional invention to be examined (37 CFR § 1.129(b))	
1801	770	2801	385	Request for Continued Examination (RCE)	
1802	900	1802	900	Request for expedited examination of a design application	

Other fee (specify) _____

*Reduced by Basic Filing Fee Paid SUBTOTAL (3)

(\$130.00)

SUBMITTED BY

Name (Print/Type)		Registration No. (Attorney/Agent)		Telephone	
Chun-Pok Leung		41,405		650-326-2400	
Signature		Date		August 2, 2004	
					

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.



PATENT
Attorney Docket No.: 16869P-079400US
Client Ref. No.: 310201611US1

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of:

MUTSUMI HOSOYA

Application No.: 10/626,049

Filed: July 23, 2003

For: HIGH-AVAILABILITY DISK
CONTROL DEVICE AND
FAILURE PROCESSING
METHOD THEREOF AND
HIGH-AVAILABILITY DISK
SUBSYSTEM

Customer No.: 20350

Examiner: Unassigned

Technology Center/Art Unit: 2186

Confirmation No.: 1442

**PETITION TO MAKE SPECIAL FOR
NEW APPLICATION UNDER M.P.E.P.
§ 708.02, VIII & 37 C.F.R. § 1.102(d)**

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

This is a petition to make special the above-identified application under MPEP § 708.02, VIII & 37 C.F.R. § 1.102(d). The application has not received any examination by an Examiner.

(a) The Commissioner is authorized to charge the petition fee of \$130 under 37 C.F.R. § 1.17(i) and any other fees associated with this paper to Deposit Account 20-1430.

08/05/2004 HVDUNG1 00000055 P01430

10626049

01 FC:1460

130.00 DA

(b) All the claims are believed to be directed to a single invention. If the Office determines that all the claims presented are not obviously directed to a single invention, then Applicant will make an election without traverse as a prerequisite to the grant of special status.

(c) Pre-examination searches were made of U.S. issued patents, including a classification search and a computer database search. The searches were performed on or around June 17, 2004. The classification search covered Classes 710 (subclass 316), 711 (subclass 113), and 714 (subclasses 5 and 6), and was conducted by a professional search firm, Kramer & Amado, P.C. The computer database search was conducted on the USPTO systems EAST and WEST. The inventor further provided two references considered most closely related to the subject matter of the present application (see references #5 and #6 below), which were cited in the Information Disclosure Statement filed with the application on July 23, 2003.

(d) The following references, copies of which are attached herewith, are deemed most closely related to the subject matter encompassed by the claims:

- (1) U.S. Patent Publication No. 2003/0084237 A1;
- (2) Japanese Patent Publication No. 9-198308;
- (3) U.S. Patent No. 5,724,542;
- (4) U.S. Patent No. 5,615,330;
- (5) Japanese Patent Publication No. 2002-041348; and
- (6) Japanese Patent Publication No. 2002-242434.

(e) Set forth below is a detailed discussion of references which points out with particularity how the claimed subject matter is distinguishable over the references.

A. Claimed Embodiments of the Present Invention

The claimed embodiments relate to a high-availability disk control device that at no time, including at times of failure, leads to performance degradation in the storage system or to malfunctions in host applications.

Independent claim 1 recites a disk control device comprising a plurality of host interface modules configured to interface with a computer; a plurality of disk interface modules configured to interface with a storage device; a plurality of cache memory modules configured to temporarily store data read from or written to the storage device; and a switch network connecting the host interface modules, the cache memory modules, and the disk interface modules, the switch network comprising at least one switch. Each of the host interface modules is configured to execute data transfers between the computer and the cache memory modules, and each of the disk interface modules is configured to execute data transfers between the storage device and the cache memory modules. Each of the host interface modules, the disk interface modules, and the cache memory modules includes identification information providing unique identification within the switch network. The switch network includes a memory containing path information based on the identification information for data transfer paths among the host interface modules, the disk interface modules, and the cache memory modules. Each of the cache memory modules is configured to monitor failure in the cache memory module and to control changing of the path information relating to the cache memory module in the memory of the switch network.

Independent claim 6 recites a disk control device comprising a plurality of host interface modules configured to interface with a computer; a plurality of disk interface modules configured to interface with a storage device; a plurality of cache memory modules configured to temporarily store data read from or written to the storage device; a plurality of resource management modules configured to store control information relating to data transfer among the cache memory modules and the host interface modules and the disk interface modules; and a switch network connecting the host interface modules, the cache memory modules, the resource management modules, and the disk interface modules, the switch network comprising at least one switch. Each of the host interface modules is

configured to execute data transfers between the computer and the cache memory modules; and each of the disk interface modules is configured to execute data transfers between the storage device and the cache memory modules. Each of the host interface modules, the disk interface modules, the resource management modules, and the cache memory modules includes identification information providing unique identification within the switch network. The switch network includes a memory containing path information based on identification information for data transfer paths among the host interface modules, the disk interface modules, the resource management modules, and the cache memory modules. Each of the resource management modules is configured to monitor failure in the resource management module and to control changing of the path information relating to the resource management module in the memory of the switch network.

Independent claim 14 recites a disk control device comprising a plurality of host interface modules configured to interface with a computer; a plurality of disk interface modules configured to interface with a storage device; a plurality of cache memory modules configured to temporarily store data read from or written to the storage device; wherein each of the host interface modules is configured to execute data transfers between the computer and the cache memory modules, and each of the disk interface modules is configured to execute data transfers between the storage device and the cache memory modules; wherein each of the host interface modules, the disk interface modules, and the cache memory modules includes identification information providing unique identification; means for connecting the host interface modules, the cache memory modules, and the disk interface modules; and means for providing a memory containing path information based on identification information for data transfer paths among the host interface modules, the disk interface modules, and the cache memory modules, and for changing the path information for the data transfer paths in the memory, when a failure takes place in one of the cache memory modules, to avoid a failed cache memory module.

Independent claim 17 recites a disk control device comprising a plurality of host interface modules configured to interface with a computer; a plurality of disk interface modules configured to interface with a storage device; a plurality of cache memory modules configured to temporarily store data read from or written to the storage device; a plurality of resource management modules configured to store control information relating to data

transfer among the cache memory modules and the host interface modules and the disk interface modules; wherein each of the host interface modules is configured to execute data transfers between the computer and the cache memory modules, and each of the disk interface modules is configured to execute data transfers between the storage device and the cache memory modules; wherein each of the host interface modules, the disk interface modules, the resource management modules, and the cache memory modules includes identification information providing unique identification; means for connecting the host interface modules, the cache memory modules, the resource management modules, and the disk interface modules; and means for providing a memory containing path information based on identification information for data transfer paths among the host interface modules, the disk interface modules, the resource management modules, and the cache memory modules, and for changing the path information for the data transfer paths in the memory, when a failure takes place in one of the cache memory modules or the resource management modules, to avoid a failed module.

Independent claim 19 recites a failure recovery processing method for a disk control device, the method comprising providing a plurality of host interface modules configured to interface with a computer; and providing a plurality of disk interface modules configured to interface with a storage device; providing a plurality of cache memory modules configured to temporarily store data read from or written to the storage device. Each of the host interface modules is configured to execute data transfers between the computer and the cache memory modules, and each of the disk interface modules is configured to execute data transfers between the storage device and the cache memory modules. Each of the host interface modules, the disk interface modules, and the cache memory modules includes identification information providing unique identification. The method further comprises connecting the host interface modules, the cache memory modules, and the disk interface modules; providing a memory containing path information based on identification information for data transfer paths among the host interface modules, the disk interface modules, and the cache memory modules; and changing the path information for the data transfer paths in the memory, when a failure takes place in one of the cache memory modules, to avoid a failed cache memory module.

Independent claim 32 recites a disk array system for connecting to a plurality of computers via a first network. The disk array system comprises a plurality of magnetic disk devices and a disk control device connected via a second network. The disk control device comprises a plurality of host interface modules including an interface with the computers; a plurality of disk interface modules including an interface with the magnetic disk devices; and a plurality of cache memory modules connected between the plurality of host interface modules and the plurality of disk interface modules via a switch network having at least one switch. The plurality of host interface modules, the plurality of disk interface modules, and the plurality of cache memory modules each include an ID providing unique identification within the switch network. The switch includes a memory containing path information based on the IDs for data transfer paths among the host interface modules, the disk interface modules, and the cache memory modules. The disk control device comprises means for changing the path information in the memory of the switch and the IDs.

Independent claim 38 recites a disk control device comprising a plurality of host interface modules configured to interface with a computer; a plurality of disk interface modules configured to interface with a storage device; a plurality of cache memory modules configured to temporarily store data read from or written to the storage device; and a switch network connecting the host interface modules, the cache memory modules, and the disk interface modules, the switch network comprising a processor and a memory storing a program executable by the processor. Each of the host interface modules is configured to execute data transfers between the computer and the cache memory modules, and each of the disk interface modules is configured to execute data transfers between the storage device and the cache memory modules. Each of the host interface modules, the disk interface modules, and the cache memory modules includes identification information providing unique identification within the switch network. The memory of the switch network includes path information based on the identification information for data transfer paths among the host interface modules, the disk interface modules, and the cache memory modules. The program in the memory of the switch network includes a code module for changing the path information relating to the cache memory modules in response to an instruction from one of the cache memory modules upon detecting failure in the cache memory module.

One benefit that may be derived is that failure notification from the failure monitoring mechanism is analyzed by the path control mechanism and a forwarding table is controlled, thereby allowing the handling of flexible system structures. In large-scale disk control devices with multiple disk control subunits, failure information from multiple failure monitoring mechanisms can be collected by the path control mechanism to provide more reliable failure status analysis, thus providing reliable failure recovery processing.

B. Discussion of the References

None of the following references disclose or suggest providing a disk control device in which each of the host interface modules, the disk interface modules, and the cache memory modules includes identification information providing unique identification; connecting the host interface modules, the cache memory modules, and the disk interface modules; providing a memory containing path information based on identification information for data transfer paths among the host interface modules, the disk interface modules, and the cache memory modules; and changing the path information for the data transfer paths in the memory, when a failure takes place in one of the cache memory modules, to avoid a failed cache memory module.

1. U.S. Patent Publication No. 2003/0084237 A1

This reference discloses a disk array controller 1 having a plurality of disk array controlling units 1-2 and a host switch interface section 30. The disk array controlling unit 1-2 is provided with an interface (a channel interface section) 11 with the host switch interface section 30, an interface section (a disc interface section) 12 with a magnetic disc unit 5, and a cache memory section 14, in which unit a mutual connection network 21 intervenes between the channel and disc interface sections 11 and 12 and the cache memory section.14. The cache memory sections 14 of the respective disk array controlling units 1-2 are interconnected through the mutual connection network 21. It is arranged such that all of the channel interface sections 11 and the disc interface sections 12 are through the mutual connection network 21 accessible to all the cache memory sections 14. The mutual connection network 21 is arranged such that the data transfer performance thereof within a

disk array controlling unit is superior to that by way of the plurality of disk array controlling units.

2. Japanese Patent Publication No. 9-198308

This reference discloses a plurality of host computers 101, 102 connected with secondary storage devices 110, 120 through the switch 103. A cache memory 107 provided with a cache controller 104 common to the secondary storage devices 110, 120 is switch-connected to the secondary storage devices in parallel. At the time of accessing data, the host computers 101, 102 transmit data to the cache memory 107 through the switch. When a cache error occurs in the cache memory 107, a disk array management table is referred to, a port to which the secondary storage device corresponding to a logic volume in a packet is connected is specified and the secondary storage device is accessed. Data transmitted from the secondary storage device is stored in the cache memory 107 and data is transferred to the host computers 101, 102 through the switch 103.

3. U.S. Patent No. 5,724,542

This reference discloses a method of controlling disk control unit. The disk control unit 30 is equipped with first and second groups each having a channel adapter 31a for interfacing the host apparatus 11; a device adapter 32a for interfacing the direct-access storage device 40a; a resource manager 35a, equipment with a control table store, for performing control related to overall resource management and processing operations, and a service adapter 36a for executing initial microprogram loading, status monitoring processing, and malfunction recovery processing; and a cache memory 33 provided so as to be shared by the first and second groups. Data commanded from the host apparatus is written in the disk device via the cache memory or data read from the disk device is transferred to the host apparatus via the cache memory.

4. U.S. Patent No. 5,615,330

This reference discloses a method for rapidly recovering a multiprocessor data processing system from failure of a boot disk. Each data processing unit 10, 11 in the system has a private boot disk (D0-1, D0-2 with private boot disk drive 14), and at least one shared

disk (D1-1, D1-2 with shared disk drive 15). If the boot disk of one of the processing units fails, the system is temporarily reconfigured to connect a new boot disk in place of the shared disk in that processing unit. Another of the processing units is then operated to copy the contents of its own boot disk to the new boot disk.

5. Japanese Patent Publication No. 2002-041348

This reference relates to a communication pass through mechanism for providing network communication with high availability between a shared system resource and a client of the system resource. The system resource is provided with a control/processing subsystem with many peer blade processors. Ports of each blade processor are connected with each client/server network path and each client is connected with corresponding ports of each blade processor. Each blade processor is provided with a network failure detector to transfer beacon transmission with other blade processors via the corresponding blade processor port and a network path. Each blade processor redirects client communication to a failed port of other blade processor to the corresponding port of the blade processor by accepting that no beacon transmission is received from a failed port of other blade processor. As with other conventional approaches, this technique involves updating routing tables for each of the multiple processors, which renders failure handling time-consuming, prevents continuation of read/write tasks from the host computer, and can lead to performance degradation in the storage system and malfunction in application programs.

6. Japanese Patent Publication No. 2000-242434

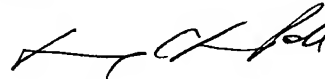
This reference discloses a storage device system constructed to correspond to the scale or request of a computer system with the goal of easily realizing the extension of a storage device system and improvement in reliability in the future. The system 1 has a plurality of subset 10 having a storage device for holding data and a controller for controlling the storage device and switch device 20 arranged between the subsets 10 and a host 30. Each switch device 20 has a managing table for holding management information for managing the configuration of the storage device system 1. According to the management information, address information contained in the frame information outputted by the host 30 is translated and the frame information is distributed to the subsets 10. In this conventional technology, a

Appl. No. 10/626,049
Petition to Make Special

failure in one of the multiple disk array subsets is handled by updating routes and the like by interpreting packets within the switch and modifying requests to the failed sections so that their destinations are changed to redundant sections having equivalent functions.

(f) In view of this petition, the Examiner is respectfully requested to issue a first Office Action at an early date.

Respectfully submitted,



Chun-Pok Leung
Reg. No. 41,405

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California 94111-3834
Tel: 650-326-2400
Fax: 415-576-0300
Attachments
RL:rl
60263047 v1

(19)



JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **09198308 A**(43) Date of publication of application: **31.07.97**

(51) Int. Cl. **G06F 12/08**
G06F 12/08
G06F 3/06
G06F 3/06

(21) Application number: **08023202**(22) Date of filing: **17.01.96**(71) Applicant: **HITACHI LTD**

(72) Inventor: **FUJIBAYASHI AKIRA**
TAKAMOTO YOSHIFUMI

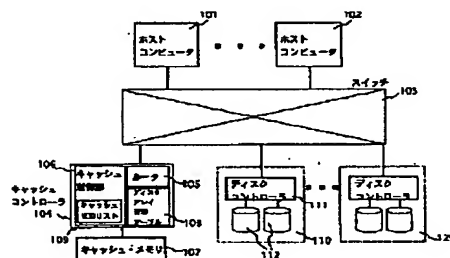
(54) **DATA STORAGE SYSTEM**

COPYRIGHT: (C)1997,JPO

(57) Abstract:

PROBLEM TO BE SOLVED: To effectively use a cache memory and to improve data access performance in a data storage system using a switch.

SOLUTION: Plural host computers 101 and 102 are connected with secondary storage devices 110 and 120 through the switch 103. A cache memory 107 provided with a cache controller 104 common to the secondary storage devices 110 and 120 is switch-connected to the secondary storage devices in parallel. At the time of accessing data, the host computers 101 and 102 transmit data to the cache memory 107 through the switch. When a cache error occurs in the cache memory 107, a disk array management table 108 is referred to, a port to which the secondary storage device corresponding to a logic volume in a packet is connected is specified and the secondary storage device is accessed. Data transmitted from the secondary storage device is stored in the cache memory 107 and data is transferred to the host computers 101 and 102 through the switch 103.



(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平9-198308

(43)公開日 平成9年(1997)7月31日

(51)Int.Cl.*	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 12/08		7623-5B	G 0 6 F 12/08	H
		7623-5B		P
	3 2 0	7623-5B		3 2 0
3/06	3 0 1		3/06	3 0 1 M
	3 0 2			3 0 2 A
審査請求 未請求 請求項の数 5 F D (全 10 頁)				

審査請求 未請求 請求項の数5 FD (全 10 頁)

(21)出願番号 特願平8-23202

(22)出題日 平成8年(1996)1月17日

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 發明者 藤林 昭

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(72)発明者 高本 良史

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

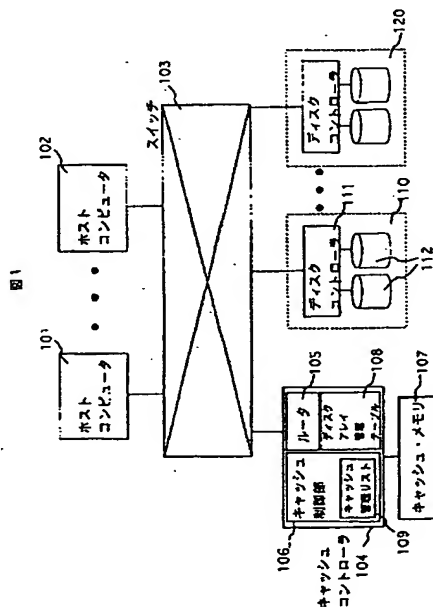
(74)代理人 弁理士 笹岡 茂 (外1名)

(54) 【発明の名称】 データ記憶システム

(57)【要約】

【課題】 スイッチを用いたデータ記憶システムにおいて、キャッシュメモリを有効に活用してデータアクセス性能を向上することにある。

【解決手段】 複数のホストコンピュータ101,102と2次記憶装置110,120はスイッチ103を介して接続され、2次記憶装置に共通のキャッシュコントローラ104を備えるキャッシュメモリ107が2次記憶装置と並列にスイッチに接続される。ホストコンピュータはデータをアクセスするときバケットをスイッチを介してキャッシュメモリに送出する。キャッシュメモリでキャッシュミスが生じたときは、ディスクアレイ管理テーブル107を参照してバケット内の論理ボリュームに対応する2次記憶装置が接続されているポートを特定して2次記憶装置をアクセスし、2次記憶装置から送られてくるデータをキャッシュメモリに格納し、データをスイッチを介してホストコンピュータに転送する。



【特許請求の範囲】

【請求項1】 複数のホストコンピュータと、複数の2次記憶装置とから構成され、該複数のホストコンピュータと該複数の2次記憶装置との間をスイッチにより接続するデータ記憶システムにおいて、該スイッチに該複数の2次記憶装置と並列に独立したキャッシュメモリを接続し、該キャッシュメモリは該複数の2次記憶装置の該スイッチを介したキャッシュメモリであることを特徴とするデータ記憶システム。

【請求項2】 請求項1記載のデータ記憶システムにおいて、前記キャッシュメモリは、前記ホストコンピュータのデータ転送要求に対して、要求されたデータを検索し、検索の結果該データがキャッシュメモリ内にない場合には、前記スイッチを介して該データが格納されている2次記憶装置から該データを取り込み、該キャッシュメモリに格納した後にホストコンピュータに該データを転送することを特徴とするデータ記憶システム。

【請求項3】 請求項2記載のデータ記憶システムにおいて、前記キャッシュメモリはキャッシュコントローラを備え、該キャッシュコントローラは、複数の各2次記憶装置が接続される前記スイッチのポート番号とホストコンピュータの認識している論理ボリュームを対応させたディスクアレイ管理テーブルを有し、前記検索の結果前記データがキャッシュメモリ内にない場合、該ディスクアレイ管理テーブルを参照して2次記憶装置をアクセスすることを特徴とするデータ記憶システム。

【請求項4】 請求項3記載のデータ記憶システムにおいて、前記各ホストコンピュータのオペレーションシステムに前記ディスクアレイ管理テーブルを設け、該各ホストコンピュータは、前記複数の2次記憶装置のいずれかに直接アクセスするとき、該ディスクアレイ管理テーブルを参照して2次記憶装置をアクセスすることを特徴とするデータ記憶システム。

【請求項5】 請求項3記載のデータ記憶システムにおいて、前記スイッチは、前記各ホストコンピュータが前記キャッシュメモリを使用するか直接前記2次記憶装置を使用するかを示すテーブルと直接前記2次記憶装置を使用する場合に用いられる前記ディスクアレイ管理テーブルと同様のテーブルからなるホスト管理テーブルを有し、前記ホストコンピュータからのアクセス要求に応じて前記ホスト管理テーブルを参照して前記キャッシュメモリまたは前記2次記憶装置に前記ホストコンピュータからのアクセスデータを転送することを特徴とするデータ記憶システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、複数のホストコンピュータと複数の2次記憶装置をスイッチを利用して接続する構成のデータ記憶システムに関する。

【0002】

【従来の技術】一般的なデータ処理システムは、ホストコンピュータと2次記憶装置から構成されている。2次記憶装置として使用されるのは主に磁気ディスク装置である。ここで発明者のいう磁気ディスク装置は単体のディスクドライブまたはディスクアレイを意味する。本発明の利用分野として挙げたスイッチを利用したデータ記憶システムの一例が特開平7-44322号において記述されている。

【0003】スイッチを用いてホストコンピュータと磁気ディスク装置を接続する構成のデータ記憶システムでは、ホストコンピュータおよび磁気ディスク装置の台数の変更にも柔軟に対応することが可能でスケラビリティに優れる。このスイッチを利用したホストコンピュータと磁気ディスク装置間のデータ転送はバケット交換によって実現される。これを簡単に説明すると、ホストコンピュータはコマンドバケット（例えば、ポート番号1番；セクタ番号2；データサイズ5.12 Byte；Readのような構成）を送信し、そのコマンドバケットをスイッチが解析してホストコンピュータと要求されたポート間の接続が確立したのちに、ホストコンピュータとそのポートに接続している磁気ディスク装置間でデータ転送を開始すると言う手順になる。また、従来のバススイッチと異なり、ある1組のホストコンピュータとの磁気ディスク装置の間でデータ転送が行われていても、データバスは占有されず、同時に他のホストコンピュータと磁気ディスク装置の間でもデータ転送が行える。ここで、データの読出しに関して、磁気ディスク装置では、ホストコンピュータから頻繁にアクセス要求のあるデータに対して、その都度読出し動作を行うと、メカニカルな動作を伴うためデータ転送に時間がかかるという問題がある。この問題を解決するために、磁気ディスク装置は通常キャッシュメモリを備えており、頻繁にアクセス要求のあるデータは、そのキャッシュメモリに格納することでそのデータの読出しに対するメカニカルな動作を省略し性能の向上を図っている。しかし、従来例の一つとして挙げたシステム構成では、従来のバススイッチを本発明の用いているスイッチに置き換え、ホストコンピュータとのインターフェース以下の部分はアレイコントローラが磁気ディスク装置とホストコンピュータ間のデータ転送を制御している従来の一般的構成である。このような構成の場合には、接続された磁気ディスク装置の各々にキャッシュメモリが分散されて配置されており、それぞれのキャッシュが有効に利用されていない。

【0004】

【発明が解決しようとする課題】従来例として挙げたデータ記憶システムにおいて、接続している複数の磁気ディ

スク装置のそれぞれにキャッシュメモリが備わっている場合、性能向上の為のキャッシュメモリが、接続している磁気ディスク装置それぞれに分散して配置されることになり、それぞれのキャッシュを有効に利用できない。例えば、ある一つの磁気ディスク装置に格納されているデータに頻繁にアクセス要求が送られてくるような場合には、他の磁気ディスク装置に備わっているキャッシュメモリはほとんど利用されないことになり無駄になってしまう。

【0005】本発明の目的は、スイッチを用いたデータ記憶システムにおいて、キャッシュメモリを有効に活用してデータアクセス性能を向上することにある。

【0006】

【課題を解決するための手段】上記目的を達成するため、本発明は、複数のホストコンピュータと、複数の2次記憶装置とから構成され、該複数のホストコンピュータと該複数の2次記憶装置との間をスイッチにより接続するデータ記憶システムにおいて、該スイッチに該複数の2次記憶装置と並列に独立したキャッシュメモリを接続し、該キャッシュメモリは該複数の2次記憶装置の該スイッチを介したキャッシュメモリであるようにしている。また、前記キャッシュメモリは、前記ホストコンピュータのデータ転送要求に対して、要求されたデータを検索し、検索の結果該データがキャッシュメモリ内がない場合には、前記スイッチを介して該データが格納されている2次記憶装置から該データを取り込み、該キャッシュメモリに格納した後にホストコンピュータに該データを転送するようにしている。前記キャッシュメモリはキャッシュコントローラを備え、該キャッシュコントローラは、複数の各2次記憶装置が接続される前記スイッチのポート番号とホストコンピュータの認識している論理ボリュームを対応させたディスクアレイ管理テーブルを有し、前記検索の結果前記データがキャッシュメモリ内がない場合、該ディスクアレイ管理テーブルを参照して2次記憶装置をアクセスするようにしている。また、前記各ホストコンピュータのオペレーションシステムに前記ディスクアレイ管理テーブルを設け、該各ホストコンピュータは、前記複数の2次記憶装置のいずれかに直接アクセスするとき、該ディスクアレイ管理テーブルを参照して2次記憶装置をアクセスするようにしている。また、前記スイッチは、前記各ホストコンピュータが前記キャッシュメモリを使用するか直接前記2次記憶装置を使用するかを示すテーブルと直接前記2次記憶装置を使用する場合に用いられる前記ディスクアレイ管理テーブルと同様のテーブルからなるホスト管理テーブルを有し、前記ホストコンピュータからのアクセス要求に応じて前記ホスト管理テーブルを参照して前記キャッシュメモリまたは前記2次記憶装置に前記ホストコンピュータからのアクセスデータを転送するようにしている。

【0007】

【実施例】本発明の提供するデータ記憶システムを以下に図面を示し実施例を参照して詳細に説明する。図1は、本発明によるデータ記憶システムの構成をブロック図で示したものである。101、102はホストコンピュータであり、103はスイッチであり、104はキャッシュコントローラであり、ルータ105とキャッシュ制御部106から構成される。107はキャッシュメモリである。108はキャッシュコントローラ104の持つディスクアレイ管理テーブルである。109はキャッシュ制御部106がキャッシュ検索時に参照するキャッシュ管理リストである。110、120は磁気ディスク装置であり、この図1中では一例としてディスクアレイ装置としており、ディスクコントローラ111、複数のディスク装置112から構成されている。磁気ディスク装置120も同様の構成である。この図に示すように本発明はキャッシュメモリ107を、複数の磁気ディスク装置110、120と共にスイッチ103に並列に接続することで、キャッシュメモリ107は前記磁気ディスク装置110、120及びホストコンピュータ101、102と、同様の1つのデバイスのごとくスイッチにより容易にアクセスできる構成となる。そして、磁気ディスク装置110、120個々にはキャッシュメモリを設けず、キャッシュメモリ107を磁気ディスク装置110、120が共用する。

【0008】ここで、スイッチ103の内部構造の一例を図7に示す。701はデータのシリアル/パラレル変換を行う部分(S/P)である。702はスイッチ制御装置である。703はスイッチ機構である。ここで、スイッチ103の動作の一例を簡単に説明する。図7の中に示すような構成のバケット704がポート1に接続されているホストコンピュータから発行される。バケットは送信先ポート番号とデータ部からなる。データ部の内容は705に示すように、コマンド、論理ボリューム(Lvol#)、ブロック番号(BLK#)、ホスト番号(Host#)からなる。送信されてきたバケットの送信先ポートをみて、スイッチ制御装置702がそのポートとの接続を確立するように制御線を通じて信号を送出する。送信先ポートとの接続が確立したのち、送信先に向けてバケットまたはデータが送られる。

【0009】次に上記システム構成においてキャッシュメモリ107を有効利用するために用いるデータ転送方法について述べる。図1のキャッシュ制御部106で行うホストコンピュータからのデータ読出し要求に対するデータ転送のフローチャートを図2に示す。まず、ホストコンピュータはデータ読出し要求をキャッシュメモリの接続ポートに対して送る。ステップ201において、要求されたデータをキャッシュ制御部106において、送信されてきたバケット内の論理ボリュームLvolとブロック番号BLKに基づいてキャッシュ管理リスト109を参照しキャッシュメモリ107内にデータが存在す

るかどうか検索する。次に判断ステップ202において、要求されたデータがキャッシュメモリ107内に存在する(キャッシュヒット)場合には、ステップ205において、要求されたデータをホストコンピュータに転送する。要求されたデータが存在しない(キャッシュミス)場合には、ステップ203において、ルータに制御を移す。続いてステップ204において、ルータの制御で磁気ディスク装置から送られて来たデータをキャッシュ制御部106がキャッシュメモリ107に格納し、キャッシュ管理リストを更新する。この時、キャッシュメモリの容量一杯までデータがすでに格納されている場合は、データの追い出しが必要となる。現在追い出しの手法は様々な方法が利用されているが、ここでは最も使用頻度が少なく最も古いデータを追い出しの候補に選ぶ方法を使用することとするが、その他の方法を用いたとしても本発明の実施にはなんら問題はない。そして、ステップ205において、ホストコンピュータに要求されたデータを転送する。

【0010】ここで、図2のステップ201においてキャッシュ制御部が参照するキャッシュ管理リストの概略を図9に示す。キャッシュコントローラ内のルータが受信したバケット内の論理ボリューム番号とブロック番号がキャッシュ制御部に渡される。キャッシュメモリ内のデータはキャッシュ管理リストによりブロック単位に管理されている。現在使用中(キャッシュメモリ内にデータが存在する)論理ボリューム番号とブロック番号のリストと未使用のリストを持っており、渡された論理ボリューム番号とブロック番号が使用リスト中に存在するかどうか検索する。

【0011】図3は前記ルータに制御を移した後の処理をフローチャートで示したものである。ここで、ルータの機能は、簡略に説明すると、ホストコンピュータから転送されたバケット内のディスクコマンドを解析し、その結果に基づいて所定の磁気ディスク装置を選択しコマンドやデータをルーティングすることである。図3中のステップ301において、ルータは送信されてきたバケット内の論理ボリュームをみて、ディスクアレイ管理テーブル(以下、図4において説明する。)を参照し、その論理ボリュームに対応する磁気ディスク装置が接続されているポートを特定する。ステップ302において、特定したポートに対しバケットを送出する。ステップ303において、このバケットを受けた磁気ディスク装置から転送されてくるデータを受け取る。

【0012】これまで述べてきた実施例の通り、ホストコンピュータへのデータ転送時は常時キャッシュメモリ107を使用することで、例えば磁気ディスク装置101に格納されているデータに頻繁にアクセスがあるような場合でも、他の磁気ディスク装置に格納されているデータもキャッシュメモリ107に格納しておくことができ、すべての磁気ディスク装置に対してキャッシュメモ

リ107が有効利用される。

【0013】図4は前記ディスクアレイ管理テーブルである。ホストコンピュータのオペレーティングシステム(以下、OSと記述する)は、この論理ボリューム(Lvol)とブロック(BLK)によりデータを指定してアクセス要求を発行する。実際のデータは複数ある磁気ディスク装置内に格納されているので、その磁気ディスク装置の接続されているポートとの対応を取るためにこのディスクアレイ管理テーブルを用いる。このディスクアレイ管理テーブルをキャッシュコントローラ104が持ち、前記実施例のようにホストコンピュータへのデータ転送時に常時キャッシュメモリ107を利用することで、ホストコンピュータのOSは論理ボリュームを用いてデータアクセス要求を発行することになり、複数の磁気ディスク装置が接続されていることとその内部のデータ配列の構成をユーザーに意識させず、単一の磁気ディスク装置の使用環境を提供する。

【0014】図5では本発明の提供するシステムにおいて、実際のデータ転送の一例を簡略に示している。ここでは、ホストコンピュータ101からデータ1への読み出し要求が発行され、ホストコンピュータ102からデータ2への読み出し要求が発行されている場合を考える。データ1は前記キャッシュメモリ内に現在格納されており、データ2はこのキャッシュメモリ内に無く磁気ディスク装置120に格納されている。ホストコンピュータ101からのデータ1の読み出し要求を図5中に示すような構成のバケット501として送信する。図2で示したフローチャートに従い、先ず最初に上記キャッシュメモリ107内を検索される。データ1はキャッシュヒットするので、キャッシュメモリ107からホストコンピュータ101に転送される。同様にホストコンピュータ102はバケット502を送信して、上記キャッシュメモリ107内を検索されるがキャッシュミスとなり、キャッシュコントローラ内のルータが、図4で説明したディスクアレイ管理テーブルを参照してデータ2の格納位置を磁気ディスク装置120の接続されているポートと特定し、これに対してデータ2の読み出し要求のバケット503を発行して、データを取り込み、上記キャッシュメモリ107内に格納してから、ホストコンピュータ102に転送する。

【0015】これまでに説明した実施例では接続するすべてのホストコンピュータは本発明の提供するキャッシュメモリにアクセスするシステムになっている。しかし、ホストコンピュータによっては、使用するアプリケーションの性質により、本発明の提供するキャッシュメモリより直接磁気ディスク装置にアクセスする方が性能が良いという場合がある。これに対して、以下に説明する実施例に対応する。一つの実施例は、ディスクアレイ管理テーブルをホストコンピュータのOSにも持たせることである。これにより、ホストコンピュータには論理

ボリュームと実際にデータの格納されている磁気ディスク装置が接続されているポートの対応が分かっているの
で、発行するバケットの送信先ポートにキャッシュメモリまたは磁気ディスク装置の接続されているポートを指定することで、本発明のキャッシュメモリの使用、不使用が選択可能になる（概略図を図6に示す。）

もう一つの実施例は前記スイッチにホストコンピュータのアクセス要求の管理を行う構成管理テーブルを持たせることである。図8にその概略図を示す。このホスト管理テーブル801は、スイッチ103内のスイッチ制御装置702に接続され、送信されてきたバケット内のホスト番号とホスト管理テーブル801を照合して、キャッシュメモリ107を使用するホストコンピュータと使用しないホストコンピュータを判断し、ホストコンピュータの発行する転送要求先をキャッシュメモリ107か、要求するデータの格納されている磁気ディスク装置かに振り分ける。例えば、ホストコンピュータの発行するアクセス要求をホスト管理テーブル801を参照してキャッシュメモリを使用しないホストである場合、ホスト管理テーブルはディスクアレイ管理テーブルと同様の

【0016】

【発明の効果】本発明により、スイッチを用いたデータ記憶システムにおいて、キャッシュメモリを有効に活用してデータアクセス性能が向上する。さらに、ホストコンピュータの使用するアプリケーションによってキャッシュメモリを使用するか、キャッシュメモリを不使用として直接磁気ディスク装置にアクセスするかの選択性を有する効率のよいデータ記憶システムを可能とする。

【図面の簡単な説明】

【図1】本発明によるキャッシュメモリを備えるスイッチを利用したデータ記憶システムの構成を示すブロック図である。

【図2】ホストコンピュータからの読出し要求に対するキャッシュ制御部の処理のフローチャートを示す図である。

【図3】ルータがデータをキャッシュメモリに格納する場合の処理のフローチャートを示す図である。

【図4】ディスクアレイ管理テーブルの一例を示す図である。

【図5】本発明における実際のデータの流れの一例を示した概略図である。

【図6】ディスクアレイ管理テーブルをホストコンピュータのOSに持たせた場合の本発明のデータ記憶システムの構成を示すブロック図である。

【図7】スイッチの構造の概略図である。

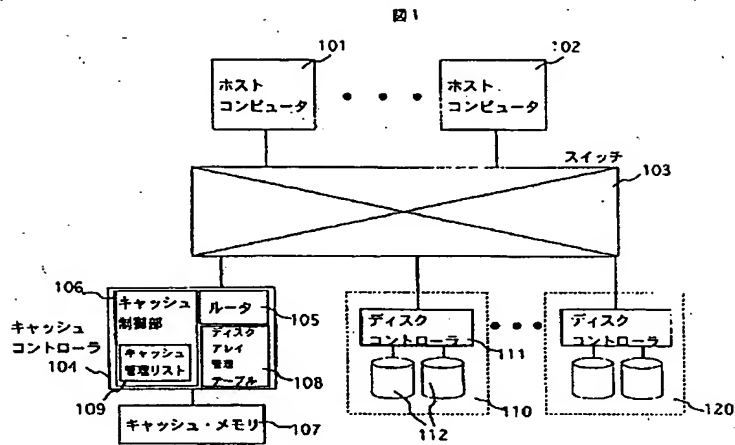
【図8】スイッチがホスト管理テーブルを備える場合の本発明のデータ記憶システムの構成を示すブロック図である。

【図9】キャッシュ管理リストの概略図である。

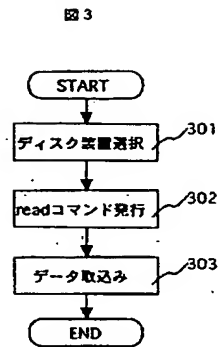
【符号の説明】

- 101、102 ホストコンピュータ
- 103 スイッチ
- 104 キャッシュコントローラ
- 105 ルータ
- 106 キャッシュ制御部
- 107 キャッシュメモリ
- 108 ディスクアレイ管理テーブル
- 110、120 磁気ディスク装置
- 111 ディスクコントローラ
- 112 ディスク装置
- 701 S/P
- 702 スイッチ制御装置
- 703 スイッチ機構
- 801 ホスト管理テーブル

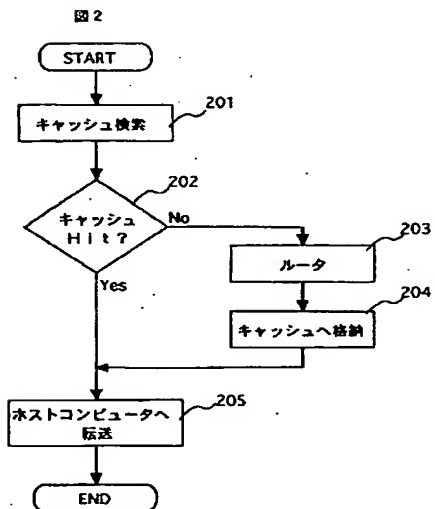
【図1】



【図3】



【図2】

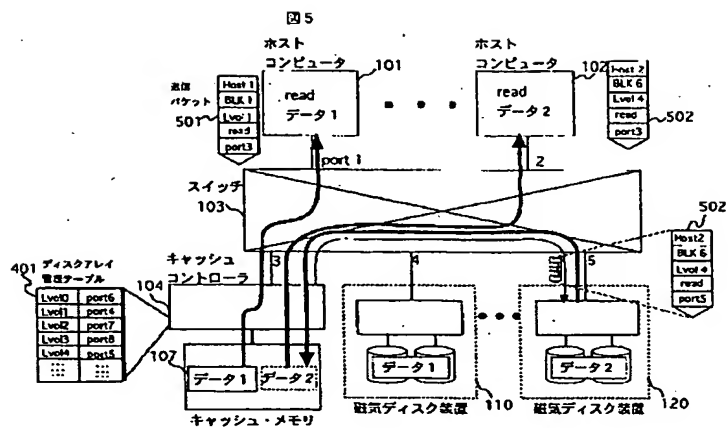


【図4】

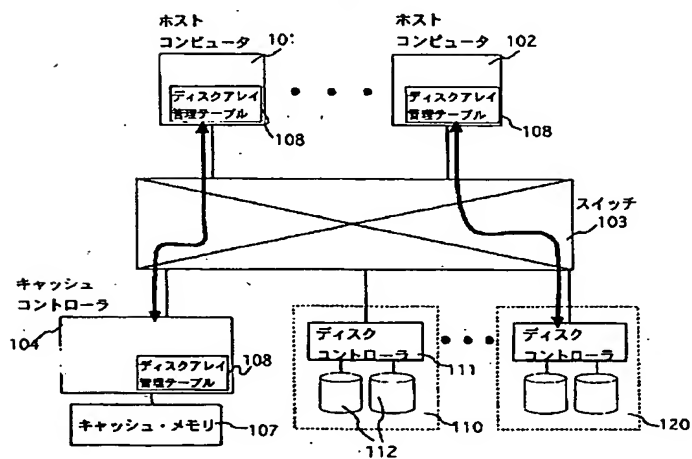
図4

管理ボリューム番号	接続先ポート番号
0	ポート1
1	ポート1
2	ポート2
3	ポート2
4	ポート3
...	...
...	...

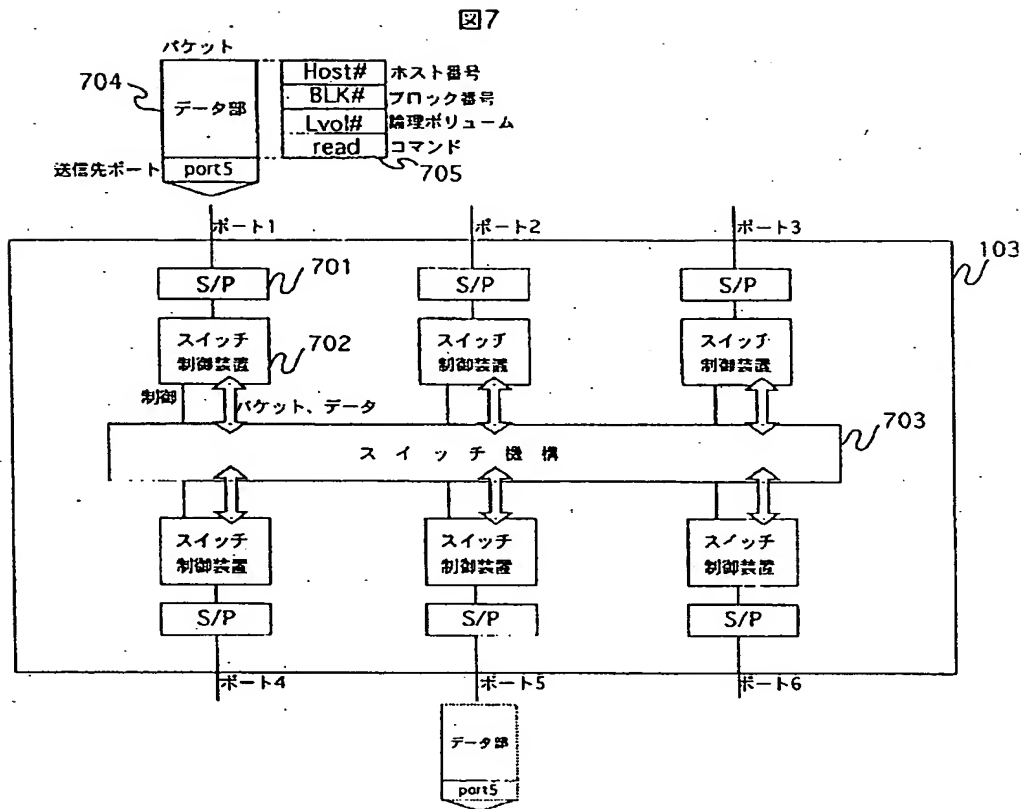
ディスクレイ管理テーブル



6

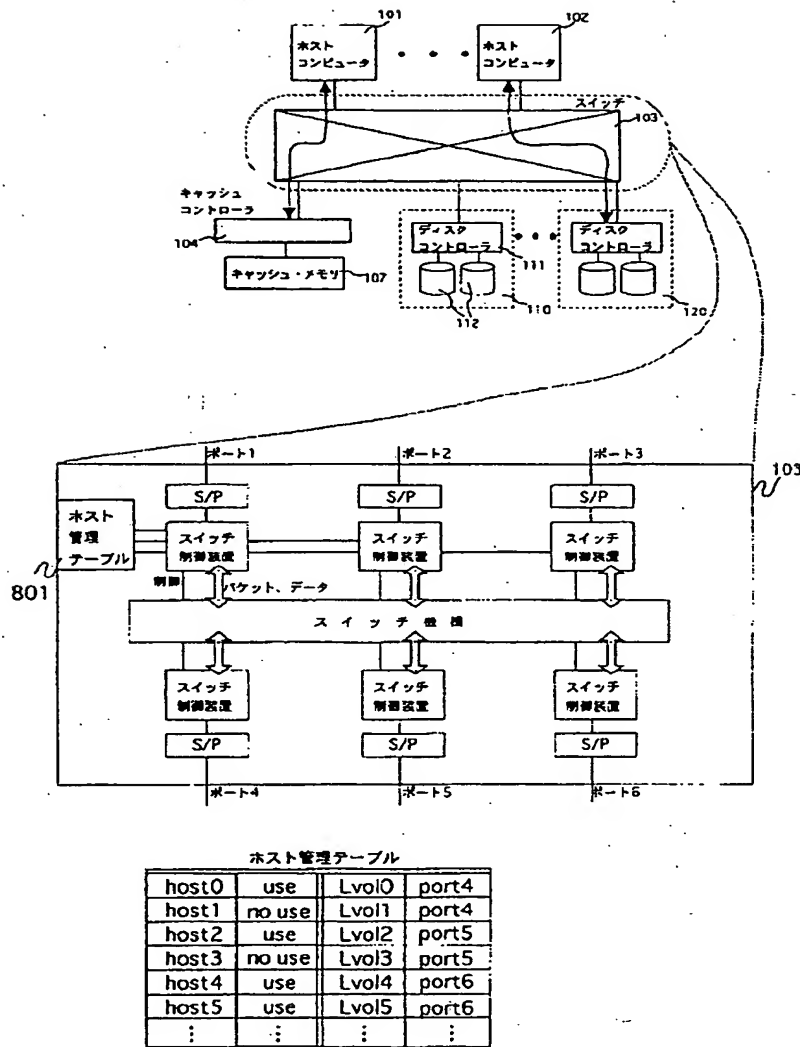


【図7】



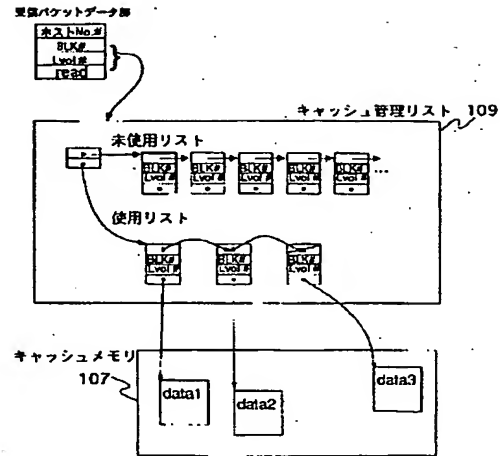
【図8】

図8



【図9】

図9



PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-041348

(43)Date of publication of application : 08.02.2002

(51)Int.Cl.

G06F 12/00

G06F 13/00

(21)Application number : 2001-155798

(71)Applicant : EMC CORP

(22)Date of filing : 24.05.2001

(72)Inventor : SCOTT JOHN A
JONES JAMES GREGORY

(30)Priority

Priority number : 2000 579428

Priority date : 26.05.2000

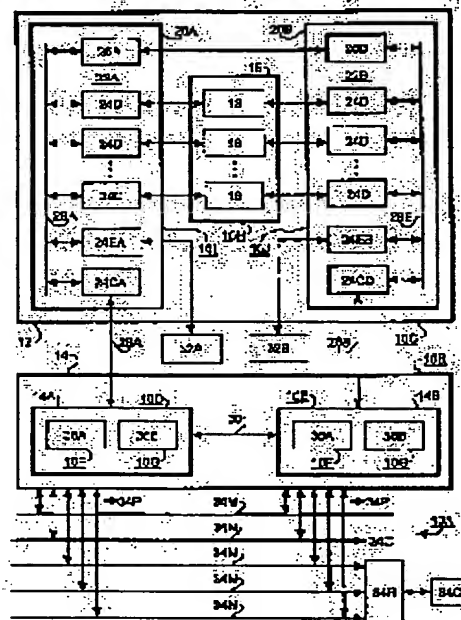
Priority country : US

(54) COMMUNICATION PASS THROUGH SHARED SYSTEM RESOURCE TO PROVIDE COMMUNICATION WITH HIGH AVAILABILITY, NETWORK FILE SERVER AND ITS METHOD

(57)Abstract:

PROBLEM TO BE SOLVED: To provide communication pass through mechanism to provide network communication with high availability between a shared system resource and a client of the system resource.

SOLUTION: The system resource is provided with a control/processing sub-system with many peer blade processors. Ports of each blade processor are connected with each client/server network path and each client is connected with corresponding ports of each blade processor. Each blade processor is provided with a network failure detector to transfer beacon transmission with other blade processors via the corresponding blade processor port and a network path. Each blade processor redirects client communication to a failed port of other blade processor to the corresponding port of the blade processor by accepting that no beacon transmission is received from a failed port of other blade processor.



[Date of registration]

[Number of appeal against examiner's decision
of rejection]

[Date of requesting appeal against examiner's
decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2002-41348

(P2002-41348A)

(43) 公開日 平成14年2月8日(2002.2.8)

(51) Int.Cl. ⁷	識別記号	F I	テマコード* (参考)
G 0 6 F 12/00	5 4 5	G 0 6 F 12/00	5 4 5 A 5 B 0 8 2
13/00	3 0 1	13/00	3 0 1 P 5 B 0 8 3
	3 5 1		3 5 1 M 5 B 0 8 9

審査請求 有 請求項の数 8 O L (全 25 頁)

(21) 出願番号 特願2001-155798(P2001-155798)
 (22) 出願日 平成13年5月24日(2001.5.24)
 (31) 優先権主張番号 09/579428
 (32) 優先日 平成12年5月26日(2000.5.26)
 (33) 優先権主張国 米国 (US)

(71) 出願人 500131642
 イーエムシー コーポレーション
 アメリカ合衆国 マサチューセッツ州
 01748 ホブキントン サウス ストリート 171
 (72) 発明者 ジョン エー スコット
 アメリカ合衆国 ノースカロライナ州
 27513 キャリー トラファルガー レーン 102
 (74) 代理人 100082500
 弁理士 足立 勉

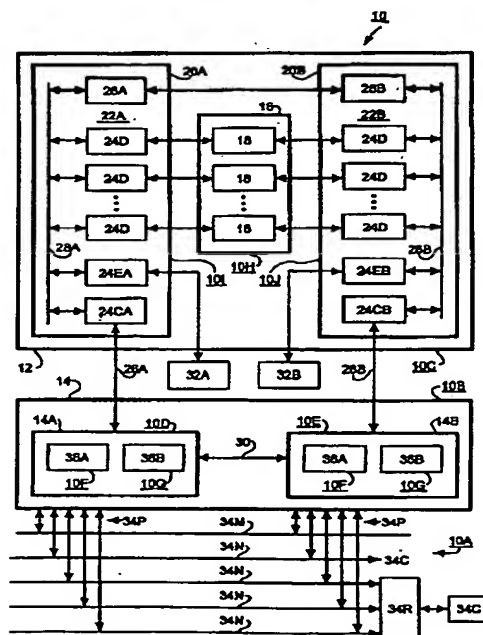
最終頁に続く

(54) 【発明の名称】 可用性が高い通信を提供する通信バススルー共有システムリソース、ネットワークファイルサーバ及び方法

(57) 【要約】

【課題】 共有システムリソースとシステムリソースのクライアントとの間で可用性の高いネットワーク通信を提供する通信バススルー機構を提供する。

【解決手段】 システムリソースは、多数のピアブレイドプロセッサを備えた制御/処理サブシステムを備える。各ブレイドプロセッサのポートは、各クライアント/サーバネットワークバスに接続され、各クライアントは、各ブレイドプロセッサの対応するポートに接続されている。各ブレイドプロセッサは、対応するブレイドプロセッサポート及びネットワークバスを介して他のブレイドプロセッサとピーコン伝送をやりとりするネットワーク故障検出器を備える。各ブレイドプロセッサは、他のブレイドプロセッサの故障したポートからピーコン伝送を受領できなかったことを受けて、他のブレイドプロセッサの故障したポートへのクライアント通信をブレイドプロセッサの対応するポートへリダイレクトする。



【特許請求の範囲】

【請求項1】 複数のクライアント/サーバ通信バスを含むネットワークを介してシステムリソースと通信するクライアントにシステムリソースサービスを提供するシステムリソースが、システムリソース操作を実行するためのシステムリソースサブシステムと、制御/処理サブシステムとを備え、制御/処理サブシステムが

多数のピアブレイドプロセッサを備え、各ブレイドプロセッサが各クライアント/サーバネットワーク通信バスに接続されたポートを備えるとともに各クライアントが各ブレイドプロセッサの対応するポートに接続され、各ブレイドプロセッサが、

各クライアントの通信ルートを決定する通信ルーティングテーブルを備えた、ブレイドプロセッサとクライアントとの間の通信操作をサポートするネットワーク機構と、

ブレイドプロセッサ及びシステムリソースサブシステム間の通信とブレイドプロセッサ間の相互プロセッサ通信リンクとを提供する相互プロセッサ通信プロセッサと、通信モニタリング機構とを備え、通信モニタリング機構が、

ブレイドプロセッサの対応するポートに接続するネットワーク通信バスを介して別のブレイドプロセッサとピーコン伝送をやりとりするためのネットワーク故障検出器と、

他のブレイドプロセッサの故障したポートからのピーコン伝送を受領できなかった際、ネットワーク故障検出器に応じて、そのブレイドプロセッサの対応するポートに、故障したポートへのクライアント通信をリダイレクトするクライアントへのリダイレクションメッセージを送信するための応答ジェネレータと、

応答ジェネレータの操作に応じて、リダイレクションメッセージに対応するように通信ルーティングテーブルを修正し、相互プロセッサ通信リンクを介して他のブレイドプロセッサとのクライアント通信をルーティングするためのバスマネージャとを備えるシステムリソース。

【請求項2】 各ブレイドプロセッサが、さらに、別のブレイドプロセッサとの相互プロセッサ通信リンクの故障を検出し、通信ルーティングテーブルを読み取ってそのブレイドプロセッサと他のブレイドプロセッサとの対応するポート間の機能するネットワーク通信バスを選択し、通信ルーティングテーブルを修正して選択された機能するネットワーク通信バスを介した相互プロセッサ通信リンクを介して相互プロセッサ通信をリダイレクトするための、相互ブレイド通信モニタを備えることを特徴とする請求項1に記載のシステムリソース。

【請求項3】 複数のクライアント/サーバ通信バスを含むネットワークを介してシステムリソース及びシステムリソースと通信するクライアント間で高可用性を備えた通信を提供する、故障に耐性がある共有システムに使用される通信バススルー機構であって、通信バススルー機構が、

システムリソース操作を実行するためのシステムリソースサブシステムと、

多数のピアブレイドプロセッサを備えた制御/処理サブシステムとを備え、各ブレイドプロセッサが各クライアント/サーバネットワーク通信バスに接続されたポートを備えるとともに各クライアントが各ブレイドプロセッサの対応するポートに接続され、

各ブレイドプロセッサが、

15 各クライアントの通信ルートを決定する通信ルーティングテーブルを備えた、ブレイドプロセッサとクライアントとの間の通信操作をサポートするネットワーク機構と、

20 ブレイドプロセッサ及びシステムリソースサブシステム間の通信とブレイドプロセッサ間の相互プロセッサ通信リンクとを提供する相互プロセッサ通信プロセッサと、通信モニタリング機構とを備え、通信モニタリング機構が、

25 ブレイドプロセッサの対応するポートに接続するネットワーク通信バスを介してブレイドプロセッサと別のブレイドプロセッサとの間でピーコン伝送をやりとりするためのネットワーク故障検出器と、

他のブレイドプロセッサの故障したポートからのピーコン伝送を受領できなかった際、ネットワーク故障検出器に応じて、そのブレイドプロセッサの対応するポートに、故障したポートへのクライアント通信をリダイレクトするクライアントへのリダイレクションメッセージを送信するための応答ジェネレータと、

35 応答ジェネレータの操作に応じて、リダイレクションメッセージに対応するように通信ルーティングテーブルを修正し、相互プロセッサ通信リンクを介して他のブレイドプロセッサとのクライアント通信をルーティングするためのバスマネージャとを備えることを特徴とする通信バススルー機構。

【請求項4】 各ブレイドプロセッサが、さらにブレイドプロセッサと別のブレイドプロセッサとの間の相互プロセッサ通信リンクの故障を検出し、

45 通信ルーティングテーブルを読み取ってそのブレイドプロセッサと他のブレイドプロセッサとの対応するポート間の機能するネットワーク通信バスを選択し、

通信ルーティングテーブルを修正して選択された機能するネットワーク通信バスを介して相互プロセッサ通信をリダイレクトするための、

相互ブレイド通信モニタを備えることを特徴とする請求項3に記載のシステムリソース。

【請求項5】複数のクライアント／サーバ通信バスを含むネットワークを介してファイルサーバ及びファイルサーバのクライアント間で高可用性を備えた通信を提供する通信バススルー機構を備えた、故障に耐性があるネットワークサーバであって、ネットワークサーバが、クライアントファイルシステム共有資源を保存するための記憶サブシステムと、

多数のブレイドプロセッサを備えた制御／処理サブシステムとを備え、各ブレイドプロセッサが各クライアント／サーバネットワーク通信バスに接続されたポートを備えるとともに各クライアントが各ブレイドプロセッサの対応するポートに接続され、

各ブレイドプロセッサが、各クライアントの通信ルートを決定する通信ルーティングテーブルを備えた、ブレイドプロセッサとクライアントとの間の通信操作をサポートするネットワーク機構と、

ブレイドプロセッサ及び記憶サブシステム間の通信とブレイドプロセッサ間の相互プロセッサ通信リンクとを提供する相互プロセッサ通信プロセッサと、通信モニタリング機構とを備え、通信モニタリング機構が、

ブレイドプロセッサの対応するポートに接続するネットワーク通信バスを介してブレイドプロセッサと別のブレイドプロセッサとの間でビーコン伝送をやりとりするためのネットワーク故障検出器と、

他のブレイドプロセッサの故障したポートからのビーコン伝送を受領できなかった際、ネットワーク故障検出器に応じて、そのブレイドプロセッサの対応するポートに、故障したポートへのクライアント通信をリダイレクトするクライアントへのリダイレクションメッセージを送信するための応答ジェネレータと、

応答ジェネレータの操作に応じて、リダイレクションメッセージに対応するように通信ルーティングテーブルを修正し、相互プロセッサ通信リンクを介して他のブレイドプロセッサとのクライアント通信をルーティングするためのバスマネージャとを備えることを特徴とするネットワークファイルサーバ。

【請求項6】 各ブレイドプロセッサが、さらに、ブレイドプロセッサと別のブレイドプロセッサとの間の相互プロセッサ通信リンクの故障を検出し、通信ルーティングテーブルを読み取ってそれらのブレイドプロセッサのポート間の機能するネットワーク通信バスを選択し、

通信ルーティングテーブルを修正して選択された機能するネットワーク通信バスを介して相互プロセッサ通信をリダイレクトするための、

相互ブレイド通信モニタを備えることを特徴とする請求項5のファイルサーバ。

【請求項7】複数のクライアント／サーバ通信バスを含

むネットワークを介してシステムリソースと通信するクライアントにシステムリソースサービスを提供するリソースシステムにおいて、システムリソースとシステムリソースのクライアントとの間で高可用性を備えた通信を提供する方法であって、システムリソースが、システム

リソース操作を実行するためのシステムリソースサブシステムと多数のブレイドプロセッサを備えた制御／処理サブシステムとを備え、各ブレイドプロセッサが各クライアント／サーバネットワーク通信バスに接続されたポートを備えるとともに各クライアントが各ブレイド

プロセッサの対応するポートに接続され、各ブレイドプロセッサが、ブレイドプロセッサとクライアントとの間の通信操作をサポートするネットワーク機構と、ブレイドプロセッサとシステムリソースサブシステムとの間の通信を提供する相互プロセッサ通信プロセッサとを備え、方法が、

ブレイドプロセッサにおいて、ブレイドプロセッサの対応するポートを接続するネットワーク通信バスを介して他のブレイドプロセッサとビーコン伝送をやりとりすることにより別のブレイドプロセッサの通信操作をモニタリングするステップと、

他のブレイドプロセッサの故障したポートからのビーコン伝送を受領できなかった際、そのブレイドプロセッサの対応するポートに、故障したポートへのクライアント通信をリダイレクトするクライアントへのリダイレクションメッセージを送信するステップと、

相互プロセッサ通信リンクを介して他のブレイドプロセッサとのリダイレクトされたクライアント通信をルーティングするステップとを備えた方法。

【請求項8】 システムリソースとシステムリソースのクライアントとの間で高可用性を備えた通信を提供する請求項7の方法であって、方法が、さらに、

ブレイドプロセッサにおいて、ブレイドプロセッサと別のブレイドプロセッサとの間の相互プロセッサ通信リンクの故障を検出するステップと、

ブレイドプロセッサと他のブレイドプロセッサとの対応するポートの間の機能するネットワーク通信バスを選択するステップと、

選択された機能するネットワーク通信バスを介して相互プロセッサ通信をリダイレクトするステップとを備えることを特徴とする方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、ネットワークファイルサーバのような、故障に耐性がありレイテンシが低い共有システムリソースにおける高レベルトランザクションロギング機構のためのシステム及び方法、特に、多重サーバシステムリソースにおいて利用されるクロスサーバ高レベルミラードトランザクションロギング機構に

関する。

【0002】

【従来の技術】コンピュータシステムにおいて絶えず問題となるのは、安全で故障に耐性があるリソースを提供すること、例えばコンピュータシステムとコンピュータシステムのクライアントまたはユーザとの間の通信が故障の際にも維持されるような通信リソース、そして故障の際にデータが失われずかつ損失を被ることなく回復または再構築されるようなデータ記憶リソースを提供することである。この問題は、システムデータ記憶機器のように、通常、共有リソースが1つ以上のシステムリソース、例えば、多数のクライアント間で共有され、システムネットワークを通じてアクセスされるファイルサーバから構成されるネットワークシステムにおいては特に解決するのが難しい。共有リソースにおける故障、例えば、ファイルサーバのデータ記憶機能における故障、あるいはファイルサーバのクライアントとファイルサーバによりサポートされるクライアントファイルシステムとの間の通信における故障は、システム全体の故障に発展する恐れがある。この問題は、データ量及び通信量と、ファイルサーバのような共有リソースによってサポートされるデータトランザクション数とが単一クライアントシステム内におけるそれらに比べて著しく大きいという点で特に厳しいものであり、その結果、リソース、データトランザクション、クライアント／サーバ通信における複雑さを著しく大きくしてしまう。この複雑さの増大は故障の可能性を増大させ、故障からの回復をより難しくする。さらに、その問題は、故障が、ディスクドライブや制御プロセッサ、あるいはネットワーク通信のような、数多くのリソースコンポーネントまたは関連する機能のどれにでも起こりうるという点で多次元的である。また、共有リソース通信及びサービスが1つ以上のコンポーネントに故障が起きても利用可能であり続け、さらに、リソースの操作が、完了した操作及びトランザクションと、故障が起きたときに実行されていた操作及びトランザクションとの両方について保存され回復されることが望ましい。

【0003】ネットワークファイルサーバシステムを従来技術の共有システムリソースの典型的な例として考えると、従来技術のファイルサーバシステムは、クライアント／サーバ通信及びファイルサーバのファイルトランザクション機能においてフォールトトレランスを達成するため、そしてデータの回復または再構築のために数多くの方法を採用してきた。これらの方法は、リダンダンシ、すなわち、複写システムエレメントの供給と、故障したエレメントの複写エレメントへの置き換え、あるいは失われた情報を再構築するのに用いられる情報の複写コピーの作成とに基づくものが代表的である。

【0004】例えば、従来技術の多くのシステムが、データ及びファイルトランザクションの保存及び回復に業

界標準のRAID技術を組み込んでいる。RAID技術は、予備のデータ及びエラー訂正情報を複数のディスクドライブの予備アレイに渡って分散する一群の方法である。故障したディスクドライブは予備のドライブに置換され、故障したディスクのデータは予備のデータ及びエラー訂正情報から再構築される。従来技術のその他のシステムは、クライアント／ファイルサーバ通信及びクライアント／クライアントファイルシステム通信の信頼性及び可用性を高めるために、故障した通信バスまたはファイルプロセッサからの通信またはファイル処理を同等の並列バスまたはプロセッサに切り換える適当なスイッチング機能を備えた多重複写式並列通信バスまたは多重複写式並列処理ユニットを採用している。しかしながら、これらの方法は、主要な通信バス及び処理バスの複写、そして、故障したエレメントを機能するエレメントに交換するのに複雑な管理及び同期機構を必要とするので、システムリソースに多額の費用がかかる。また、これらの方法により、故障の際にサービス及び機能が継続して実行され、例えばRAIDの利用により、完了したデータトランザクション、すなわち、ディスク上の固定記憶装置にコミットされたトランザクションが回復または再構築されるが、これらの方法は、トランザクションの実行中の故障により失われたトランザクションの再構築または回復をサポートしない。

【0005】この結果、従来技術の別の方法においては、トランザクションの実行中に起きる故障により失われたトランザクションの回復及び再構築のために情報リダンダンシが利用される。これらの方法には、キャッシング、トランザクションロギング、ミラーリングが含まれる。キャッシングとは、固定記憶装置、すなわちディスクドライブへのデータの移動により固定記憶装置にデータトランザクションがコミットされるまで、あるいはデータトランザクションが固定記憶装置から読み取られて受け手に送られるまで、固定記憶装置への及びそれからのデータフローバスのメモリ中にデータを一時的に記憶することである。トランザクションロギング、あるいはジャーナリングとは、データトランザクションが固定記憶装置にコミットされるまで、すなわちファイルサーバにおいて完了されるまで、一時的にデータトランザクションを記述する情報、すなわち要求されたファイルサーバ操作を記憶し、さらに、記憶された情報から失われたデータトランザクションを再構築または再実行することである。ミラーリングは、多くの場合キャッシングまたはトランザクションロギングと共に用いられ、基本的に、キャッシュまたはトランザクションログの記録がファイルプロセッサで生成されるときに、例えば、別のプロセッサのメモリまたは固定記憶空間にキャッシュまたはトランザクションログの内容のコピーを保存することである。

【0006】しかしながら、キャッシング、トランザク

ションロギング、ミラーリングは、あまり満足のいくものではない。なぜなら、それらは多くの場合システムリソースを高額にし、キャッシング、トランザクションロギング、ミラーリング機能及びそれに続くトランザクションの回復操作を行うために複雑な管理及同期操作と、機構とを必要とし、著しくファイルサーバのレイテンシ、すなわちファイルトランザクションを完了するのに要する時間を増加するからである。また、キャッシング及びトランザクションロギングは、キャッシング及びロギング機構が存在するプロセッサの故障に弱いこと、また、ミラーリングがキャッシュまたはトランザクションログの内容の損失問題への解決である一方で、ミラーリングは、キャッシングまたはトランザクションロギングと同様の欠点を有することに注意しなければならない。これらの問題は、キャッシングと、特にトランザクションロギング及びミラーリングとがトランザクションロギングの間に莫大な量の情報の保存を必要とする点、及び、ログファイルトランザクションの再構築または再実行が、ファイルトランザクションの再構築のために、トランザクションログを分析し、再生し、ロールバックする複雑なアルゴリズムの実装を必要とする点でより複雑となる。また、これらの方法が、各データトランザクションが非常に多くの詳細で複雑なファイルシステム操作として実行されているようなより低いレベルのファイルサーバ機能で実装される場合が多いという点で、これらの問題はさらに複雑になる。その結果、抽出され保存されるべき情報量と、データあるいはデータトランザクションを抽出して保存し、データまたはデータトランザクションを回復及び再構築するために必要となる操作の数及び複雑さとは著しく増大する。

【0007】また、これらの方法はシステムリソースを割高にし、それらの方法を管理するための複雑な管理及同期機構を必要とする。そして、システムリソースが割高であるために、これらの方法が提供できるリダンダンスの度合いは制限されるので、システムは、多くの場合、複数のソースに起こる故障に対応できない。例えば、システムがある機能のために複写式並列プロセッサユニットまたは通信バスを設けても、両方のプロセッサユニットまたは通信バスで故障が起きればシステム全体が失われてしまう。さらに、通信及びデータの保存及び回復を保証するこれらの従来技術は、通常、互いから隔絶された状態で、そして異なるレベルまたはサブシステムで動作する。このため、通常、これらの方法は協力してまたは連動して動作するわけではなく、互いに相反して動作するかもしれず、複数の故障または連動した故障、またはいくつかの方法を組み合わせる必要のある故障に対応できない。従来技術のいくつかのシステムは、この問題を解決しようと努力しているが、それには、中央統一的な調整機構、またはサブシステムと、協調操作を行い、故障を扱う機構間の衝突を避けるため

の互いに関連する複雑な管理及同期機構を必要とし、そのためにまたシステムリソースにお金がかかるとともに、それ自体が故障の原因となる。

【0008】

- 05 【発明が解決しようとする課題】本発明の目的は、これらの、そしてその他の従来技術に関連する問題への解決を提供することである。本発明は、複数のクライアント／サーバ通信バスを含むネットワークを介して、システムリソースとシステムリソースのクライアントとの間に
- 10 可用性の高い通信を提供するための、故障に耐性のある共有システムリソース、例えばネットワークファイルサーバ、に使用される通信バススルー機構及び通信バススルー機構の操作方法に関する。

【0009】

- 15 【課題を解決するための手段及び発明の効果】本発明によると、システムリソースには、システムリソース操作を実行するためのシステムリソースサブシステムと、多数のピアブレイドプロセッサを備えた制御／処理サブシステムとが含まれる。各ブレイドプロセッサは、各クライアント／サーバネットワーク通信バスに接続されたポートを備え、各クライアントは、各ブレイドプロセッサの対応するポートに接続されている。各ブレイドプロセッサは、各クライアントの通信ルートを決定する通信ルーティングテーブルを備えた、ブレイドプロセッサとク
- 20 ライアントとの間の通信操作をサポートするネットワーク機構と、ブレイドプロセッサとシステムリソースサブシステムとの間で通信を提供する相互プロセッサ通信プロセッサと、ブレイドプロセッサ間の相互プロセッサ通信リンクとを備える。各ブレイドプロセッサは、さらに、ブレイドプロセッサの対応するポートを接続するネットワーク通信バスを介して、別のブレイドプロセッサとピーコン伝送をやりとりするためのネットワーク故障検出器を備えた通信モニタリング機構を備える。各ブレイドプロセッサは、他のブレイドプロセッサの故障した
- 25 ポートからピーコン伝送を受領できなかった際、ネットワーク故障検出器に応じて、ブレイドプロセッサの対応するポートへ反対側のブレイドプロセッサの故障したポートへのクライアント通信をリダイレクトするクライアントへのリダイレクションメッセージを送信するための
- 30 応答ジェネレータを備える。ブレイドプロセッサのバスマネージャは、応答ジェネレータの操作に応じて、リダイレクションメッセージに対応するように通信ルーティングテーブルを修正し、相互プロセッサ通信リンクを介して他のブレイドプロセッサへ他のブレイドプロセッサ
- 35 の故障したポートへのクライアント通信をルーティングする。

- 【0010】本発明のさらなる実施例において、各ブレイドプロセッサは、相互ブレイド通信モニタを備え、別のブレイドプロセッサとの相互プロセッサ通信リンクの
- 40 故障を検出し、通信ルーティングテーブルを読み取って

ブレイドポート間の機能するネットワーク通信バスを選択し、通信ルーティングテーブルを修正して相互プロセッサ通信リンクからの相互プロセッサ通信を選択された機能するネットワーク通信バスへリダイレクトする。

【0011】

【発明の実施の形態】本発明の前述及びその他の目的、特徴、利点を、添付の図を参照しながら、実施例を用いて以下に説明する。

A. 高可用性を備えた共有リソースの概略説明 (図1)

1. 序論

以下に記述するように、本発明は、ネットワークシステムにおいて多数のユーザ間で共有されるファイルサーバ、通信サーバ、あるいはプリンタサーバのように、可用性の高いリソースに関するものである。本発明のリソースは、統合された協働クラスタからなる階層及びピアドメインから構成される。各ドメインは、リソースによってサポートされた機能またはサービスに不可欠な1つ以上の関連した機能を実行あるいは提供する。1つのドメインは、複数のサブドメインから構成されてもよいし、あるいは複数のサブドメインを具備していてもよい。例えば、1つ以上のドメインが、リソースとネットワーククライアントとの間で通信サービスを提供し、その他のドメインが、高レベルファイルシステム、通信、または印刷機能を実行し、その一方で、別のドメインが低レベルファイルシステム、通信及びプリント機能を実行してもよい。階層的に関連したドメインの場合、1つのドメインが別のドメインを制御するか、または、関連したより高いあるいは低いレベルの機能を実行することにより、より高いあるいは低いレベルのドメインをサポートすることができる。例えば、より高レベルのドメインは、関連した低レベルドメインがより低レベルのファイルまたは通信機能を実行する間、高レベルのファイルまたは通信機能を実行することができる。ピアドメインは、例えばタスクの負荷を分担してある機能についてのリソース容量を増やすために、同一あるいは並列の機能を実行したり、あるいは、共に1つのドメインを構成するために中立的なサポート関係で関連するタスクまたは機能を実行することができる。さらに、他のドメインは、ある機能についてはピアドメインであったり、他の機能については階層的に関連したドメインであったりもできる。最後に、以下に説明するように、あるドメインは、他のドメインの故障処理機構とは別に独立して動作するが、高レベルのリソース可用性を達成するために協調的に動作する故障処理機構を備える。

【0012】本発明は、例えば、そして以下に説明する目的で、高可用性を備えたネットワークファイルサーバ (HANファイルサーバ) 10 に実装される。この実装の形態を、本発明の実施例として以下に詳細に記述する。図1に示すように、本発明が実装されているHANファイルサーバ10には、例えば、データジェネラルコ

ーポレーション (Data General Corporation) のCLARiON™ファイルサーバを使用する。CLARiON™ファイルサーバは、高い可用性を備えたファイルシステム共有資源、すなわち、記憶空間をネットワーククライアントに提供するとともに、ジャーナルファイルシステム、ネットワークフェイルオーバー能力、データのバックエンドレイド (RAID) 記憶装置を利用して、共有資源に書き込まれたデータに高い整合性を提供する。本実装においては、HANファイルサーバ10は、業界標準の共通インターネットファイルシステムプロトコル (CIFS) とネットワークファイルシステム (NFS) 共有資源との両方をサポートしており、CIFS及びNFSによって使用されるようなファイルアクセス制御のための対照モデルが外からはわからないように実装されている。HANファイルサーバ10はまた、マイクロソフトウィンドズNT環境におけるドメインコントローラあるいはUNIX (登録商標) 環境のためのネットワークファイルシステム (NFS) などの既存の業界標準管理データベースを統合している。

【0013】本実装は、ゼロコピーIPプロトコルスタックを利用して高いパフォーマンスを提供する。そのために、ファイルシステムキャッシング方式をバックエンドRAID機構と緊密に統合するとともに、保存用のディスクへの書き込みを廃すために、ピア記憶プロセッサ上でミラーリングすることにより重要なデータの可用性を提供できるデュアル記憶プロセッサを使用する。以下に詳細に説明するように、本実装のHANファイルサーバ10は、デュアルプロセッサファンクショナルマルチプロセッシングモードで動作している。このモードでは、1つのプロセッサが、クライアントとディスクに存在するファイルシステムとの間でデータを転送するための全てのネットワーク及びファイルシステム操作を実行するフロントエンドプロセッサとして働き、ネットワークスタック、CIFS/NFSの実装、ジャーナルファイルシステムをサポートする。第二プロセッサは、ブロック記憶プロセッサとして働き、可用性の高いRAID構成において管理されたひとまとまりのディスクへの及びそれからのデータの読み取り及び書き込みの全ての機能を実行する。

【0014】本実装において、ファイルシステムは、カーネルベースのCIFSネットワークスタックを備えたジャーナル機能付きクイックリカバリファイルシステムとして実装され、第二モードでNFS操作をサポートするが、本発明によると、ファイルシステムのデータへのアクセスに高い可用性を提供するために修正を加えられている。ファイルシステムはさらに、ある記憶プロセッサ上のメモリに記憶されたデータ変更がその記憶プロセッサのハードウェアまたはソフトウェア故障の際に保存されるというデータ反映機能を使って、ネットワークク

クライアントがファイルシステムに加える全てのデータ変更を記憶することにより記憶プロセッサの損失に対する保護を提供する。ファイルシステムに対するコア内部のデータ変更の反映は、相互記憶プロセッサ通信システムを通じて達成され、これにより、一方の記憶プロセッサ上でクライアントによって NFS または CIFS を使用して伝達されたファイルシステムへのデータ変更は、データを記憶しているネットワーククライアントに通知が返される前に、他方の記憶プロセッサにより反映され、受領確認される。このことは、最初の記憶プロセッサ上での故障の際に代わりの記憶プロセッサにデータ変更のコピーが取り込まれ、万が一故障が起きた際には、ファイルシステムが代わりの記憶プロセッサに引き継がれた後に、その変更がファイルシステムに適用されることを保証する。後述するように、この反映機構が、ファイルを追跡するために用いられるシステムメタデータを回復及び修復する基本的なファイルシステム回復機構の頂点に構築される一方で、反映機構はユーザデータを回復あるいは修復する機構を提供する。ブロック記憶サブシステムは、RAID 技術を使用してディスクユニットの損失に対しディスクレベルでの保護を提供する。ディスクドライブが失われると、RAID 機構は、代わりのドライブにデータを再構築する機構を提供し、失われたドライブなしで動作する際、そのデータへのアクセスを提供する。

【0015】後述するように、本実装の HAN ファイルサーバ 10 は、サーバのクライアントと、予備のコンポーネント及びデータバスを利用してサーバ上でサポートされたクライアントファイルシステムとの間で可用性の高い通信を提供し、クライアントとクライアントファイルシステムとの間の通信を維持するための通信故障処理機構を提供する。本発明の HAN ファイルサーバ 10 はまた、ファイルトランザクション及びデータのバックアップ及び回復システムを備え、ファイルトランザクション及びデータの損失を防ぐとともに、ファイルトランザクション及びデータの回復または再構築を許容する。システムハードウェアまたはソフトウェア故障の際には、システムの生き残ったコンポーネントが故障したコンポーネントのタスクを引き継ぐ。例えば、記憶プロセッサ上のイーサネット（登録商標）ポートが 1 つ失われると、そのポートからのネットワークトラフィックは代わりの記憶プロセッサの別のポートによって引き継がれる。同様に、記憶プロセッサのどの部分かにその処理機能を危うくするような故障が起きたならば、全てのネットワークトラフィック及びファイルシステムが生き残った記憶プロセッサへ移転される。さらなる例では、データ及びファイルトランザクション及びバックアップ機構は、故障したコンポーネントが回復した際、故障したコンポーネントによる、あるいは対応するコンポーネントによるデータ及びファイルトランザクションの回復及び

再構築を可能にするとともに、生き残ったコンポーネントが故障したコンポーネントのファイルトランザクションを引き継ぐことを可能にする。さらに、ディスクドライブが 1 つ失われても、そのディスクのデータへのアクセスが失われない。なぜなら、RAID 機構が生き残ったディスクを用いて、失われたドライブ上にあった再構築されたデータへのアクセスを提供するからである。全てのファイルサーバに影響を及ぼす停電の際には、停電の際のファイルサーバ状態が保存され、コア内部のデータは固定記憶装置にコミットされて電源が復旧すると回復される。これにより、停電前になされた全てのデータ変更が保存される。最後に、HAN ファイルサーバ 10 の通信そしてデータ及びファイルトランザクションの故障回復機構は、サーバの各ドメインまたはサブシステムに設けられ、互いに別々に独立して機能するが、ファイルシステム通信へのクライアントの可用性を高レベルに保ち、データ及びファイルトランザクションの損失を防いで回復を可能にするために、協調的に動作する。それにも関わらず、HAN ファイルサーバ 10 の故障回復機構は、故障のソースを特定して隔離するのに通常必要な複雑な機構や手続き、さらには衝突する可能性のある故障管理操作を調整し、同期させ、管理するのに通常必要な複雑な機構及び操作を必要としない。

【0016】2. HAN ファイルサーバ 10 の詳細説明 (図 1)

図 1 には、データジェネラルコーポレーションの CLARIONTM ファイルサーバのような、本発明が実装される典型的な HAN ファイルサーバ 10 が示されている。図に示すように、HAN ファイルサーバ 10 は、記憶サブシステム 12 と、記憶サブシステム 12 を共有するデュアルコンピュータブレイド（ブレイド）14A 及び 14B からなる制御/プロセッササブシステム 14 とを備える。コンピュータブレイド 14A 及び 14B は、HAN ファイルサーバ 10 のクライアントに、ネットワークアクセス及びファイルシステム機能を提供及びサポートするために独立して動作し、相互バックアップと、ネットワークアクセス及び互いのファイルシステム機能のサポートとを提供するために協調的に動作する。

【0017】a. 記憶サブシステム 12 (図 1)

記憶サブシステム 12 は、複数のハードディスクドライブ 18 からなるドライブバンク 16 を備える。各ディスクドライブ 18 は、記憶ループモジュール 20A 及び 20B として示されるデュアル記憶ループモジュール 20 (20A 及び 20B を総称して 20 ともいう。以下同じ。)を通して双方向に読み取り/書き込みアクセスされる。図に示すように、記憶ループモジュール 20A 及び 20B にはそれぞれ、MUXBANK 22A 及び 22B として示されるマルチプレクサバンク (MUXBANK) 22 が含まれる。MUXBANK 22A 及び 22B にはそれぞれ、複数のマルチプレクサ (MUX) 24

と、ループコントローラ26A及び26Bとして示されるループコントローラ26とが含まれる。各ループコントローラモジュール20のMUX24とループコントローラ26とは、MUXループバス28A及び28Bとして示されたMUXループバス28を介して双方向に相互接続されている。

【0018】図に示すように、MUXBANK22A及び22Bにはそれぞれ、対応するディスクドライブ18に対応して接続されているディスクドライブMUX24(MUX24D)が含まれる。そのため、ドライブバンク16の各ディスクドライブ18は、MUXBANK22A及び22Bのそれぞれにおいて、対応するDMUX24Dに接続され、双方向に読み取り/書き込みされる。MUXBANK22A及び22Bはさらに、それぞれ、対応するコンピュータブレイド14A及び14Bの一方と、MUX24CA及びMUX24CBそれぞれを介して双方向に接続されており、コンピュータブレイド14A及び14Bはブレイドバス30を介して双方向に接続されている。さらに、MUXBANK22A及び22Bは、それぞれ、MUX24EA及び24EBで示される外部ディスクアレイMUX24を備えていてもよい。外部ディスクアレイMUX24は、対応するMUXループバス28A及び28Bから双方向に接続され、外部ディスクアレイMUX(EDISKA)32に双方向に接続されている。外部ディスクアレイMUX32は、図において、それぞれEDISKA32A及び32Bとして示され、予備のあるいは代替りのディスク記憶空間を提供する。

【0019】従って、各ディスクドライブ18は、MUXBANK22AのMUX24及びMUXBANK22BのMUX24と双方向に通信する。そしてMUXBANK22AのMUX24が、ループバス26Aを介して相互接続されている一方で、MUXBANK22BのMUX24は、ループバス26Bを介して相互接続されている。そのため、各ディスクドライブ18は、ループバス26A及びループバス26Bの両方を介してアクセス可能である。さらに、プロセッサブレイド14Aがループバス26Aと双方向に通信する一方で、プロセッサブレイド14Bはループバス26Bと双方向に通信する。プロセッサブレイド14A及び14Bは、直接相互接続され、ブレイドループ(ブレイド)バス30を介して通信する。このため、プロセッサブレイド14A及び14Bは、対応するループバス26を介して直接、または他方のプロセッサブレイド14を介して間接的に、どのディスクドライブ18とも双方向に通信できるとともに、相互に直接通信できる。

【0020】最後に、記憶サブシステム12について、本実施例のHANファイルサーバ10においては、例えば、各ディスクドライブ18は、簡単にユーザが置換できるキャリアに入れられたホットスワップファイバチャ

ネルディスクドライブであり、ドライブ及びキャリアは、電気を供給し、MUXループバス26A及び26Bを含む中央平面にプラグ接続される。これにより、各デュアルポートドライブをMUX24に、そしてMUX24をループコントローラ26と相互接続することができる。MUX24はファイバチャネルMUXデバイスであり、ループコントローラ26は、各MUXデバイスのバス選択を制御するマイクロコントローラを備え、各ディスクドライブ18のデュアルポートのファイバチャネルMUXループバス26A及び26Bとの接続の実行又は解除を選択的に行う。MUX24CA及び24CB、MUX24EA及び24Eは同様に、ファイバチャネルMUXデバイスであり、記憶サブシステム12をファイバチャネルループバスを介してコンピュータブレイド14A及び14BとEDISKA32A及び32Bとに接続する。コンピュータブレイドバス30も同様にファイバチャネルバスである。

【0021】b. 制御/プロセッササブシステム14(図1及び2)

前述のように、制御/プロセッササブシステム14は、コンピュータブレイドバス30を介して相互接続されるデュアルコンピュータブレイド(ブレイド)14A及び14Bからなる。コンピュータブレイド14A及び14Bは、共有記憶サブシステム12の操作を制御する計算及び制御用のサブシステムを併せ持つ。コンピュータブレイド14A及び14Bは、HANファイルサーバ10のクライアントにネットワークアクセスとファイルシステム機能とを独立して提供及びサポートし、相互バックアップと互いのネットワーク34アクセス及びファイルシステム機能のためのサポートとを協調的に提供する。図1及び2に示すように、各ブレイド14はネットワーク34に接続された多数のネットワークポート(ポート)34Pを備える。ネットワーク34は、HANファイルサーバ10とHANファイルサーバ10のクライアント34Cとの間の双方向データ通信接続を構成する。図に示すように、ネットワークには、例えば、クライアント34Cに接続する複数のクライアントネットワーク34Nと管理ネットワーク34Mとが含まれ、さらにリモートクライアント34Cに接続するルータ34Rを含むこともできる。当業者には理解されるように、ネットワーク34は、例えば、ローカルエリアネットワーク(LAN)、広域ネットワーク(WAN)、直接プロセッサ接続またはバス、ファイバオプティックリンク、あるいは前記の組み合わせから構成することができる。

【0022】図2に示すように、各ブレイド14は、メモリへの、そして通信コンポーネントのような他のエレメントへの緊密なアクセスを共有するデュアル処理ユニット36A及び36Bから構成される。各処理ユニット36A及び36Bは、フルオペレーティングシステムカーネルを実行する十分に機能的な計算処理ユニットであ

り、ファンクショナルマルチプロセッシング構造において協働する。例えば、後述されるような実装においては、一方の処理ユニット36がRAID機能を実行し、他方の処理ユニット36はネットワーク機能、プロトコルスタック機能、CIFS及びNSF機能、ファイルシステム機能を実行する。

【0023】c. HANファイルサーバ10の全体的なアーキテクチャ及びHANファイルサーバ10の故障処理機構(図1及び2)

上述のように、本発明のHANファイルサーバ10は階層及びピアドメインの集まり、すなわちノードあるいはサブシステムから構成され、各ドメインはファイルサーバの1つ以上のタスクまたは機能を実行するとともに故障処理機構を備えている。例えば、HANファイルサーバ10は、それぞれ、ネットワーク34N、制御/プロセッササブシステム14、記憶サブシステム12を有する3つの階層ドメイン10A、10B、10Cから構成され、ファイルサーバの独立した及び相補的な機能を実行する。つまり、ドメイン10Aは、クライアント34とHANファイルサーバ10との間のクライアント/サーバ通信を提供し、ドメイン10B、すなわち、制御/プロセッササブシステム14は、ドメイン10Aのクライアント/サーバ通信をサポートするとともに高レベルファイルシステムトランザクションをサポートし、ドメイン10C、すなわち、記憶サブシステム12は、クライアントのファイルシステムをサポートする。制御/プロセッササブシステム14は、2つのピアドメイン10D及び10E、すなわち、ブレイド14A及び14Bからなり、並列機能、特にクライアント/サーバ通信機能及びより高い及び低いレベルのファイルシステム操作を実行し、それにより、クライアント通信及びファイル操作のタスクの負荷を分担する。後に詳細に説明されるように、ブレイド14A及び14Bを備えたドメインはまた、クライアント/サーバ通信、ブレイド14の相互通信、高レベルファイルシステム機能、記憶サブシステム12で実行される低レベルファイルシステム機能の故障処理及びサポートを提供する独立して機能する故障処理機構を備える。各ブレイド14は、処理ユニット36A及び36Bに基づく2つの階層ドメイン10F及び10Gから構成されるドメインであり、ブレイド14A及び14Bの機能を併せ持つ別個ではあるものの相補的な機能を実行する。後述するように、一方の処理ユニット36は、高レベルファイル操作及びクライアント/サーバ通信を両機能のための故障処理機構に提供する上層ドメイン10Fを形成する。他方の処理ユニット36は、低レベルファイル操作及びブレイド14の相互通信を提供する下層ドメイン10Gを形成し、両機能及び上層ドメイン10Fのサーバ機能と故障処理機構とをサポートする独立して機能する故障処理機構を備える。最後に、記憶サブシステム12は、同様に、ディスクドライブ1

8、すなわち、サーバの記憶エレメントを構成して、ブレイド14のドメイン10EによりサポートされるRAID機構を間接的にサポートする下層ドメイン10Hと、ドメイン10D及び10Eとドメイン10Hとの間の通信をサポートする記憶ループモジュール20A及び20Bを備えたピア上層ドメイン10I及び10Jとから構成される。

【0024】従って、以下に記述するように、各HANファイルサーバ10ドメインは、1つの中央統一機構あるいは調整機構なしに、互いに独立して別々に、しかしながら互いに協調的に動作する1つ以上の故障処理機構を直接あるいは間接的に有するまたは備える。そのため、あるドメインのコンポーネントの機能あるいは操作が故障しても、関連するドメインの対応するコンポーネントが後を引き継ぐ。さらに、以下に記述するように、HANファイルサーバ10の故障処理機構は、一箇所あるいは複数箇所に故障が起きても継続した機能を提供できるように、複数の異なる技術あるいは方法を外からはわからないように採用している。

【0025】HANファイルサーバ10の全体構造及び操作をこれまで説明してきたが、以下には、HANファイルサーバ10の各ドメインをさらに詳細に、そしてHANファイルサーバ10の故障処理機構の構造及び操作を説明する。

1. ブレイド14の処理と制御コア

図2に、本実装のブレイド14を示す。ブレイド14は、デュアル処理ユニット36A及び36Bの計算コアをそれぞれ形成するプロセッサ38A及び38Bと、メモリコントローラハブ(MCH)38C、メモリ38D、入出力コントローラハブ(ICH)38Eのような多数の共有エレメントとを備える。本実装において、例えば、プロセッサ38A及び38Bは、それぞれ、内蔵のレベル2キャッシュを有するインテルペンティアムI I Iであり、MCH38C及びICH38Eはインテル820チップセットであり、メモリ38DはRDRAMあるいはSDRAMの512MB以上からなる。

【0026】図に示すように、プロセッサ38A及び38Bは、パイプラインフロントサイドバス(FSB)38F及びMCH38Cの対応するFSBポート38CAを介してMCH38Cと相互接続されている。当業者には理解されるように、MCH38C及びMCH38CのFSBポートは、プロセッサ38A及び38Bからのメモリ参照の初期化及び受信と、プロセッサ38A及び38Bからの入出力(I/O)及びメモリマップI/O要求の初期化及び受信と、メモリ38Cからプロセッサ38A及び38Bへのメモリデータの受け渡しと、メモリI/O要求から生じるメモリスヌープサイクルの初期化とをサポートする。さらに、MCH38Cはメモリ38Dへのメモリポート38CBと、ICH38Eへのハブリンクバス38Gに接続するハブリンクポート38CC

と、業界標準パーソナルコンピュータ相互接続（PCI）バスとして機能する4つのAGPポート38CDとを備えている。各PCIバスは、インテル21154チップのようなプロセッササブプロセッサブリッジユニット（P-Pブリッジ）38Hへのプロセッサに接続されている。

【0027】ICH38Eは、MCH38Cへのハブリングバス38Gに接続するハブリックポート38EA、ファームウェアメモリ38Iに接続するファームウェアポート38EB、ハードウェアモニタ（HM）38Jに接続するモニタポート38EC、ブートドライブ38Kに接続するIDEドライブポート38ED、スーパーI/Oデバイス（スーパーI/O）38Lに接続するI/Oポート38EE、他のエレメントと共に、VGAデバイス（VGA）38M及び管理ローカルエリアネットワークデバイス（LAN）38Nに接続するPCIポート38EFを含んでいる。当業者には上記の説明で十分理解されるであろう。

【0028】2. ブレイド14のパーソナルコンピュータ互換サブシステム

ICH38E、スーパーI/O38L、VGA38Mは併せてパーソナルコンピュータ（PC）互換サブシステムを構成し、ローカル制御及び表示の目的でHANファイルサーバ10のためのPC機能及びサービスを提供する。この目的のために、当業者には理解されるように、ICH38Eは、IDEコントローラ機能、IO APIC、82C59ベースのタイマ及びリアルタイムクロックを備える。スーパーI/O38Lは、例えば、標準マイクロシステムデバイスLPC47B27xであってもよく、8042キーボード/マウスコントローラ、2.88MBスーパーI/Oフロッピーディスクコントローラ、フル機能デュアルシリアルポートを提供する。一方、VGA38Mは、例えば、1MBフレームバッファメモリをサポートするシーラスロジック（Cirrus Logic）64ビットビジュアルメディア（VisualMedia）アクセラレータCL-GD5446-QCであってもよい。

【0029】3. ブレイド14のファームウェア及びBIOSサブシステム

ICH38E及びファームウェアメモリ38Iは、併せて、通常のファームウェア及びBIOS機能を実行するファームウェア及びBIOSサブシステムを構成し、その機能には、ブレイド14A及び14Bリソースのパワーオンセルフテスト（POST）及びフル設定が含まれる。例えば、AMI/Phoenixから利用できるような標準BIOSであるファームウェア及びBIOSは、1MBのフラッシュメモリを備えたファームウェアメモリ38Iに存在する。POSTが完了すると、BIOSは上述したPCIバスをスキャンし、このスキャンの間、上述及び後述する2つのPCIツーPCIブリッ

ジを設定し、以下に記述するバックエンド及びフロントエンドPCIバス上のファイバチャネル及びLANコントローラが存在を検出してPCIアドレス空間にマップする。この情報は、PC互換IOやメモリサイズなどの他の標準サイジング情報とともにIOサブシステムのトポロジを記述するMP対応テーブルに書き留められ、POSTは単純なバスチェックとメモリ診断とを実行する。POSTの終了後、綿密なりブート診断パッケージを含むフラッシュレジデントユーザバイナリコードセグメントがロードされる。リポート診断パッケージはまた、ファイバチャネルデバイスを初期化するとともに、パターンセンシティブデータを使ってデータバス及びDRAMセルを試験することにより、コンピュータブレイド上のコンポーネントの整合性をチェックする。診断が実行されると、制御はBIOSあるいはブートストラップユーティリティに戻される。制御がBIOSに移される場合は、システムはブートを続け、制御がブートストラップユーティリティに移される場合は、ブートブロックがファイバディスクから読み取られ、制御は新しくロードされたオペレーティングシステムのイメージに引き渡される。さらに、このサブシステムは、全体のシステム管理アーキテクチャをサポートする、エラーチェックロジック、環境モニタリング、エラー及びスレッショルドロギングなどの機能を提供する。最下層レベルでは、内蔵プロセッサキャッシュパリティ/ECCエラー、PCIバスパリティエラー、RDRAM ECCエラー、フロントサイドバスECCエラーを含むハードウェアエラー及び環境スレッショルドチェックが実行される。エラー及び超過の環境スレッショルドイベントは、DMI互換レコードフォーマットでフラッシュプロムの一部にロギングされる。

【0030】4. ブレイド14のI/Oバスサブシステム

最後に、MCH38C及びICH38Eは、ブレイド14の2つの入出力（I/O）バスサブシステムをサポートする。うち一方はMCH38Cによってサポートされるバックエンドバスサブシステム（BE Bus Sys）38Oであり、前述のブレイド14及び記憶サブシステム12の対応するループバス26間の双方向接続と、コンピュータブレイドバス30を介したブレイド14A及び14B間の双方向接続とを提供する。他方はICH38Eによってサポートされるフロントエンドバスサブシステム（FE Bus Sys）38Pであり、前述のネットワーク34への及びそれからの双方向接続を提供する。ネットワーク34は、前述のように、例えば、ローカルエリアネットワーク（LAN）、広域ネットワーク（WAN）、直接プロセッサ接続またはバス、ファイバオプティックリンク、あるいは前記の組み合わせであることができる。

【0031】まず、BE Bus Sys 38Oについて

考えると、上述のように、MCH38Cは、業界標準パーソナルコンピュータ相互接続(PCI)バスとして機能する4つのAGPポート38CDをサポートする。各AGPポート38CDは、インテル21154チップのようなプロセッサツープロセッサブリッジユニット(P-Pブリッジ)38Hに接続される。P-Pブリッジ38Hは、例えば、タックライト(Tach Lite)ファイバチャネルコントローラから構成される2つのファイバチャネルコントローラ(FCC)38Qの双方向バスポートに接続される。FCC38Qの並列ファイバチャネルインターフェイスは、2つの対応するシリアルライザ/デシリアルライザデバイス(SER-DES)38Rの並列ファイバチャネルインターフェイスに接続されている。一方のSER-DES38Rのシリアルインターフェイスはコンピュータブレイドバス30に接続され、他方のデュアルブレイド14への通信接続を提供する。他方のSER-DES38Rのシリアルインターフェイスは記憶サブシステム12の対応するループバス26に接続されている。

【0032】FE Bus Sys 38Pでは、上述のように、ICH38EがPCIポート38EFを備えており、図に示すように、PCIポート38EFは、PCIバスツープCIバスブリッジユニット(P-Pブリッジ)38Hと双方向に接続されている。P-Pブリッジ38Hは、例えば、双方向32ビット33MHzフロントエンドPCIバスセグメントをサポートするインテル21152から構成される。フロントエンドPCIバスセグメントは、ネットワーク34に接続する1群の双方向ネットワークデバイス(NETDEV)38Tに接続されていて、NETDEV38Tは、例えば、インテル82559 10/100イーサネットコントローラデバイスである。前述のように、ネットワーク34は、例えば、ローカルエリアネットワーク(LAN)、広域ネットワーク(WAN)、直接プロセッサ接続またはバス、ファイバオプティックリンク、あるいは前記の組み合わせであることができ、NETDEV38Tはそれに応じて選択されることが理解されるであろう。

【0033】最後に、BE Bus Sys 38O及びFE Bus Sys 38Pについて、本実施例においては、BE Bus Sys 38O及びFE Bus Sys 38Pの両方がPCIタイプのバスであり、そのため、共通の割り込み構造を有している。このため、BE Bus Sys 38O及びFE Bus Sys 38PのPCI割り込みは、BE Bus Sys 38OのPCIバスデバイスがFE Bus Sys 38PのPCIバスデバイスと割り込みを共有しないようにルーティングされる。

【0034】c. HANファイルサーバ10の操作(図1、2、3)

1. HANファイルシステム10の全体的な操作

上述のように、HANファイルシステム10は、デュアルコンピュータブレイド14を備え、各コンピュータブレイド14は記憶サブシステム12の全てのディスクドライブ18への完全なアクセスと、全てのクライアントネットワーク34Nへの接続とを有し、それぞれ独立してHANファイルシステム10の全ての機能及び操作を実行できる。ブレイド14の機能及び操作構造の概略図を図3に示す。図3は、ブレイド14A及び14Bのうちの一方を示し、他方のブレイド14は図のブレイド14と同一であり、かつミラーイメージであることが理解されるだろう。

【0035】ブレイド14の内部では、上述のように、デュアル処理ユニット36A及び36Bが、例えば、メモリコントローラハブ(MCH)38C、メモリ38D、入出力コントローラハブ(ICH)38Eのような、多数のブレイド14エレメントを共有している。処理ユニット36A及び36Bはそれぞれ、互いに独立しながらも協調的に動作し、それぞれがメモリ38Aに存在するリアルタイムオペレーティングシステム(OS)40の別々のコピーを実行する。OS40の各コピーは、例えば、処理ユニット36A及び36Bの対応する一方のために、基本メモリ管理、タスクスケジューリング、同期機能、他の基本オペレーティングシステム機能を提供する。処理ユニット36A及び36Bは、共有メモリ38Aに設けられたメッセージパッシング機構(メッセージ)42を介して通信し、メッセージは、例えば、I/Oの開始、I/Oの終了、ディスク故障のようなイベント通知、ステータスクエリー、ブレイドバス30を介してミラーリングされる、ファイルシステムジャーナルのような重要なデータ構造のミラーリングのために規定される。初期設定時、各ブレイド14はOS40と、RAIDファイルシステム及びネットワークイメージとの両方のコピーをバックエンドディスクドライブ18からロードする。それぞれ処理ユニット36A及び36Bの一方を実行する2つのRAIDカーネルは、その後、OS40の2つのインスタンス間でブレイド14のメモリ38Aを協力して分割し、OS40カーネルのコピーがロードされた後、処理ユニット36A及び36Bの操作を開始する。初期設定の後、OS40カーネルはメッセージ42を介して通信する。

【0036】図3に示すように、各ブレイド14の内部で、処理ユニット36A及び36Bの一方はバックエンドプロセッサ(BEP)44Bと称されて動作する。そして、上述のように、RAID設定ディスクへの及びそれからのデータの書き込み及び読み出しのためのブロック記憶システムとして動作するとともに、RAID機構(RAID)46を備える。RAID46には、RAIDデータ記憶及びバックアップ機能を実行するRAIDファイル機構(RAIDF)46Fと、RAID関連のシステムモニタリング機能及び以下に示す他の機能を実

行するRAIDモニタ機構(RAIDM)46Mとが含まれる。処理ユニット36A及び36Bの他方はフロントエンドプロセッサ(FEP)44Fと称されて動作し、クライアントとディスクレジデントブロック記憶システムとの間でデータを移動するための全てのネットワーク及びファイルシステム操作、そして、ネットワークドライバ、CIFS及びNFSプロトコルを含むプロトコルスタックのサポートとジャーナルファイルシステムの維持とを含めたBEP44Bの対応するRAID機能を実行する。

【0037】ブロック記憶システム操作に加えて、BEP44Bの機能には、RAIDF46F及びRAIDM46Mを介してのコアRAIDファイルシステムサポートアルゴリズムの実行、ディスクドライブ18の操作のモニタリング、自身が存在するブレイド14及びピアブレイド14の両方の操作及び状態のモニタリング、管理機能への故障の連絡が含まれる。図2及びBE Bus Sys 380について上述したように、BEP44Bはまた、BE Bus Sys 380とブレイドバス30とを介してブレイド14A及び14B間の通信を、そしてBE Bus Sys 380と記憶サブシステム12の対応するループバス26とを介してディスクドライブ18との通信をサポートする。RAIDM46Mはまた、ブレイド14の電源装置をモニタし、停電の際には適切な処理を実行する。例えば、ディスクドライブ18に重要なデータ構造の緊急書き込みを行ったり、処理ユニット36A及び36Bの生き残った方が適切な処理を開始できるように処理ユニット36A及び36Bの一方に通知をする。BEP44Bはさらに、確実なブートストラップサポート機能を提供し、それによりランタイムカーネルがディスクドライブ18に保存され、システムブートの際ロードされることができる。

【0038】FEP44Fは、ブレイド14の全てのネットワーク34関連機能及び操作を実行するネットワーク機構(ネットワーク)48を備え、FE Bus Sys 38P及びNet Dev 38Tの要素を含んでいる。例えば、ネットワーク48は、FE Bus Sys 38Pを含むネットワーククライアントに利用可能なリソースを管理及び提供し、ネットワーク34を介してクライアント34CにHANファイルシステム10へのアクセスを提供する。後述するように、ネットワーク48はまた、FEP44Fに存在する通信フェイルオーバー機構と、ここに記載されるその他の高可用性機能とをサポートする。

【0039】FEP44Fはまた、ジャーナルファイルシステム(JFile)50を含む。ジャーナルファイルシステム(JFile)50は、ネットワーク48を介してHANファイルシステム10のクライアントと、そしてメッセージ42を介してRAIDM46FのRAIDファイルシステム機能と通信する。図に示すよう

に、JFile50は、JFile50のファイルシステム機能を実行するファイルシステム機構(FSM)50Fと、FSM50Fと相互作用してそれぞれデータトランザクションのデータ及び操作をキャッシュし、データトランザクションのジャーナルを維持する内蔵書き込みキャッシュ(WCache)50C及びトランザクションログ(ログ)50Lとを含む。ログ50Lには、要求されたデータトランザクションを表すログエントリ(SE)50Eを生成するためのログジェネレータ(Log Gen)50Gと、SE50Eを記憶するログメモリ(LogM)50Mとが含まれる。LogM50Mの大きさは、以下に記述されるように、ジャーナルされるべきデータトランザクションの数に依存する。図に示すように、BEP44Bには、WCache50Cと通信して、WCache50Cの中身をミラーリングするキャッシュミラー機構(CMirror)54Mが含まれる。さらに、各ブレイド14のログ50Lは、反対側のピアブレイド14に存在するログ50Lのミラー機構(LMirror)54Lによってミラーリングされ、各ブレイド14のログ50Lは、メッセージ42、BE Bus Sys 380、ブレイドバス30を含むバスを介して対応するLMirror54Lと通信する。

【0040】最後に、FEP44Fには、ステータスモニタ機構(モニタ)52が含まれる。モニタ52は、HANファイルシステム10の変更に関するBEP44Bからの通知をモニタし、その変更を受けて適切な処理を開始する。この通知には、例えば、RAIDグループに新しく挿入されたディスクのバインディングに関する、あるいは故障したディスクのためのSNMPトラップを起動するRAIDM46Mからの通知が含まれ、モニタ52により開始される操作には、例えば、以下に記述するように、RAID機能が非常に重大なエラーに遭遇した場合等には、HANファイルサーバ10の故障処理機構によりフェイルオーバー動作を開始すること、あるいはブレイド14を完全にシャットダウンすることが含まれる。

【0041】2. HANファイルサーバ10のファイルシステム機構の操作(図1、2、3)

上記及び図3に示したように、HANファイルサーバ10のファイルサーバ機構は、3つの主要なコンポーネントあるいは層を含む。1つ目の最上層は、ブレイド14A及び14Bそれぞれのフロントエンドプロセッサ44Fに存在するWCache50C及びLog50Lを含むJFile50のファイルシステム機構である。最下層には、ディスクドライブ18を備えた記憶サブシステム12と、ブレイド14A及び14BそれぞれのBEP44Bに存在するブロック記憶システム機能及びRAIDF46F機能とが含まれる。HANファイルサーバ10ファイルシステム機構の3番目の層あるいはコンポーネントは、ファイルシステム機構の操作に影響する故障

を検出して処理し、ファイルシステム故障からの回復を行う故障処理機構から構成される。上層及び下層ファイルシステムエレメントの構造及び操作はすでに上述されており既知のエレメントと類似しているため、本実施例のHANファイルサーバ10ファイル機構のこれらのエレメントは、本発明を完全に理解するのに必要でない限りここでは詳細に説明されない。以下の記述は、その代わりに、HANファイルサーバ10ファイル機構の故障処理機構、特にHANファイルサーバ10の上層レベルのファイルシステムエレメントの操作に関する故障処理機構に焦点をあてる。

【0042】上述のように、HANファイルサーバ10ファイル機構の第3のコンポーネントは、HANファイルサーバ10コンポーネントの損失から生じるデータの損失に対する保護を提供するミラーリング機構から構成される。図3に示すように、ミラーリング機構には、各ブレイド14毎に、ブレイド14のBEP44Bに存在するキャッシュミラー機構(CMirror)54Mと、反対側のピアブレイド14のBEP44Bに存在するログミラー機構(LMirror)54Lとが含まれる。CMirror54Mは、メッセージ42を介してJFile50のWCACHE50Cと通信する継続動作キャッシュミラーリング機構である。ログ50Lは、ピアブレイド14のBEP44Bに存在するLMirror54Lにより要求に応じてミラーリングされ、メッセージ42、BE Bus Sys 380、コンピュータブレイドバス30を介して対応するLogM50Mと通信する。これにより、クライアントに通知される前に、ブレイド14Aあるいは14Bの一方を介したファイルシステムへの全データ変更が、ブレイド14Aあるいは14Bの他方に反映される。これに関連して、本実施例においては、ログ50Lのミラーリングは、各ファイルシステムトランザクションの処理中に実行される。そのため、トランザクションログミラーリングのレイテンシは実際のファイルシステムトランザクションの実行により限度ぎりぎりまで掩蔽される。最後に、RAID F46Fによりサポートされ提供されるディスクドライブ18ファイルシステム、制御、モニタリング、データ回復/再構築機能は、HANファイルサーバ10データ保護機構の一部でもあり、記憶サブシステム12内部へのデータミラーリング法を使用していることが理解されるだろう。

【0043】以下に記述されるように、これらのミラーリング機構は、よって、故障のタイプによって、ブレイド14における故障を処理する数多くの代替法をサポートしている。例えば、ブレイド14の一方が故障した際、生き残ったブレイド14は、そのLMirror54Lに保存されたファイルトランザクションを読み取り、故障したブレイド14が復帰したときに故障してい

たブレイド14に戻す。その際には、復帰したブレイド14により失われたファイルトランザクションが再実行され回復される。他の方法では、ブレイド14のネットワーク34フェイルオーバー機構について以下に記述するように、故障したブレイド14あてのファイルトランザクションが、ブレイド14間のブレイドバス30のバスを介して、あるいはブレイド14のネットワーク34フェイルオーバー機構によって生き残ったブレイド14へのクライアントのリダイレクションにより、生き残っているブレイド14にリダイレクトされる。生き残ったブレイド14は、それにより、故障したブレイド14あてのファイルトランザクションの実行を引き継ぐ。以下に記述するように、生き残ったブレイド14は、この操作の一部として、そのLMirror54Lに保存されている故障したブレイド14からのファイルトランザクションを再実行することにより故障したブレイド14の失われたファイルトランザクションを再実行して回復するか、あるいは、故障したブレイド14が復帰した後に故障していたブレイド14にファイルトランザクションを読み戻す。これにより、故障の際の故障したブレイド14上のファイルシステムの状態が再構築され、確認済みのトランザクションのために、故障したブレイドからデータが失われることはない。

【0044】3. HANファイルサーバ10の通信機構の操作 (図1、2、3)

図1、2、3に示すように、本発明に組み込まれているHANファイルサーバ10の通信機構は、3つのレベルあるいは層の通信機構から構成されるとみなすことができる。説明のために、最上層レベルは、クライアント34Cと、HANファイルサーバ10によってサポートされるクライアントファイルシステム構造との間のファイルトランザクション通信のためのネットワーク34関連通信機構、及び、関連する通信故障処理機構から構成される。通信機構の中間層には、ブレイドバス30及びメッセージ42を介したブレイド14A及び14B間の通信をサポートする通信機構と、関連する通信故障処理機構とが含まれる。通信機構の最下層には、ブレイド14及び記憶サブシステム12間、そして記憶サブシステム12のエレメント間の通信バス及び機構とが含まれる。前記は、すでに説明されており、本発明を理解するために必要でない限りさらには説明されない。

【0045】まず、HANファイルサーバ10の通信機構の上層レベルについて考える。図3に示すように、ブレイド14A及び14BそれぞれのFEP44Fに存在するネットワーク機構(ネットワーク)48は、TCP/IPプロトコルスタック(TCP/IPスタック)58を含むネットワークスタックオペレーティングシステム(NetSOS)56とネットワークデバイスドライバ(NetDD)60とを含み、以下に記述するように、これらの機構には、単一ポート34Pの故障、ネッ

トワーク34の故障、ブレイド14全体の故障を調整して処理する機能が含まれる。これに関連して、本文の他の箇所にも記載するように、ネットワーク34は、例えば、ローカルエリアネットワーク（LAN）、広域ネットワーク（WAN）、直接プロセッサ接続またはバス、ファイバオプティックリンク、あるいは前記の組み合わせから構成されることができ、NETDEV38T及びNetDD60はそれに応じて実装される。

【0046】また、図3に示され、HANファイルサーバ10の通信機構の高可用性について以下に説明されるように、各ネットワーク48はさらに、クライアントルーティングテーブル（CRT）48Aを含む。CRT48Aは、ブレイド14によりサポートされるクライアント34Cに付随するルーティング及びアドレス情報を含むクライアントルーティングエントリ（CRE）48Eと、反対側のピアブレイド14によってサポートされるクライアント34CのCRE48Eとを保存する。当業者には理解されるように、CRE48Eは、ネットワーク48によって、所定のクライアント34Cへファイルトランザクション通信を送るために利用されることができ、必要であるならば、ブレイド14に割り当てられたクライアント34Cから受領したファイルトランザクション通信を識別、あるいは確認するために利用されることもできる。図に示すように、各ネットワーク48にはまた、ブレイドルーティングテーブル（BRT）48Bが含まれる。BRT48Bは、ブレイド14にアクセス可能でブレイド14によって共有されるネットワーク34通信バスに関するアドレス及びルーティング情報を含み、これにより、ブレイド14間の利用可能な通信バスを形成する。典型的な本実装のネットワーク48において、CRT48A及びBRT48B情報は、ブレイドバス30を含む通信バスを介してブレイド14A及び14B間で通信されるが、例えば、ネットワーク34Mを介して各ブレイド14に提供されることもできる。

【0047】HANファイルサーバ10のネットワーク34通信機構の全体的な操作を説明する。図1及び2を見ると、HANファイルサーバ10の各ブレイド14は、ネットワーク34と接続して通信する複数のポート34Pをサポートしている。例えば、本実装において、各ブレイド14は合計5つのポート34Pをサポートしていて、うち4つのポート34Pはネットワーク34Nに接続されてクライアント34Cにサービスを提供し、1つのポートは、HANファイルサーバ10の管理のために予約されて管理ネットワーク34Mに接続されている。図に示すように、ブレイド14A及び14Bそれぞれの対応するポート34Pは同じネットワーク34に接続されており、そのため、各ネットワーク34は、対応するポート34Pを介して、ブレイド14A及び14Bそれぞれに接続される。本実施例において、HANファイルサーバ10のポート34Pは、10個の異なるIP

アドレス、すなわち、各ポートにつき1アドレスを設定され、ブレイド14のそれぞれ対応する組み合わせのポート34Pのポート34Pが同じネットワーク34に接続されている。そのため、各ネットワーク34は、2つのアドレス、すなわちブレイド14A及び14Bそれぞれの一方へのアドレスを介してHANファイルサーバ10をアドレス指定することができる。HANファイルサーバ10の各クライアントが割り当てられるポート34Pは、従来技術であり当業者には簡単に理解されるように、クライアントに存在するARPテーブルにより各クライアント内で決定される。さらに、図2に示すように、クライアント34Cは、HANファイルサーバ10がデフォルトのルートを設定されるかまたはRIPまたはOSPのようなルーティングプロトコルを備える場合、直接接続されたネットワーク34通信のうちの一方を介して、あるいは任意のルータ34Rを介して、HANファイルサーバ10にアクセスできる。HANファイルサーバ10の別の実装では、各クライアント34Cは、複数のネットワーク34を介してHANファイルサーバ10のポート34Pに接続されることができ、ネットワーク34は、以下に記述するように、クライアント34CのARPテーブル及びHANファイルサーバ10を適切に改良することにより、ローカルエリアネットワーク（LAN）、広域ネットワーク（WAN）、直接プロセッサ接続またはバス、ファイバオプティックリンク、あるいは前記の組み合わせのような異なる技術を利用することができる。

【0048】図3に示すように、ブレイド14A及び14Bそれぞれの各FEP44Fに存在するネットワーク48機構はさらに、CIFS62及びNFS64ネットワークファイルシステムと、その他の必要なサービスとを備える。図3には示されていないこれらの付加的なサービスには、以下のものが含まれる。

【0049】NETBIOS - リモートリソースにアクセスするためにPCクライアントによって使用されるマイクロソフト/IBM/インテルプロトコル。このプロトコルの重要な特徴の1つは、サーバ名をトランスポートアドレスに変更することであり、サーバは、共有資源、すなわち、¥server¥shareを識別するためにクライアントにより用いられるUNC名のコンポーネントとなる。HANファイルサーバ10では、サーバはブレイド14Aまたは14Bを表す。NETBIOSはまた、CIFS62パケットフレーミングを提供し、HANファイルサーバ10はRFC1001及びRFC1002に規定されるようなTCP/IPに優先してNETBIOSを使用する。

【0050】SNMP - Simple Network Management Protocol。HANファイルサーバ10に、エージェントと呼ばれる処理を提供する。エージェントは、システムについての情報

を提供するとともに、通常でないイベントが起きた際、トラップを送信する機能を提供する。

【0051】SMTP - Simple Mail Transport Protocol。通常でないイベントが起きた際、電子メールメッセージを送信するためにHANファイルサーバ10により用いられる。

NFS - サンマイクロシステムズネットワーク情報サービス。NFSファイルシステムへのアクセス制御に用いられるユーザIDを識別するためにNFSサーバによって用いられるプロトコルを提供する。

【0052】RIP - 動的ルーティングプロトコル。ルータ34Rのようなルータの背後で動作しているクライアントのサポートによりネットワークポロジを明らかにするために使用される。本実装のHANファイルサーバ10においては、このプロトコルは、ルーティング情報のモニタのために受動モードで動作する。別の実装においては、ユーザがシステム初期設定の間にデフォルトルートを設定または明示してもよい。

【0053】本発明の説明では、HANファイルサーバ10の正常動作時は、各ネットワーク48のエレメント、すなわち、NetSOS56、TCP/IPスタック58、NetDD60、CRT48Aは、クライアント34CとHANファイルサーバ10との間のネットワーク通信操作を実行するのに当業者には明らかな従来方法で動作することが当業者には理解されるであろう。このため、HANファイルサーバ10のこれらの機能についてはこれ以上説明をしない。以下はHANファイルサーバ10のネットワーク関連通信機構の高可用性に焦点をあてて説明する。

【0054】4. HANファイルサーバ10の通信故障処理機構(図1、2、3)

a. ネットワーク通信故障機構

通信または接続故障が簡単に検出される一方、どのコンポーネントが故障したのかを見極め、どんな訂正手段をとるのが適当かを判断することが難しくかつ複雑であることは当業者には明白に理解されることであろう。例えば、故障の可能性のあるソースには、ポート34P、あるいはポート34Pとネットワーク34のハブまたはスイッチとの間のリンク、あるいはブレイド14間のネットワークのパーティションが含まれるがこれに限定されるわけではない。しかしながら、HANファイルサーバ10は、ブレイド14故障と同様に、1つ以上のネットワーク34インターフェイス故障及び、異なるタイプのネットワーク34故障とに対処できるIPネットワーク通信サービスを提供し、さらに、さまざまな故障を徐々に減少させる機能をサーバシステムに提供するために、異なるクラスあるいはタイプの故障を処理する多数の協調的あるいは補足的な機構を実装する。例えば、ブレイド14のポート34Pインターフェイス故障の際、HANファイルサーバ10は、ブレイド14A及び14B間

のコンピュータブレイドバス30接続を利用して、ネットワークトラフィックをピアブレイド14上の機能している対応ポート34Pからポート34Pが故障したブレイド14へ転送することができる。この機能により、1つのネットワークポート34Pの故障によりブレイド14全体が動かなくなるのが防がれ、その結果、ブレイド14によってサポートされるファイルシステムを移動する必要がなくなる。この機能はまた、故障が異なるネットワーク34上で起きる限り、すなわち、故障がブレイド14上の対応するポート34Pの両方に起きない限り、片方あるいは両方のブレイド14上での複数のネットワークポート34P故障を調整できることが明らかである。各ネットワーク34のブレイド14の一方で少なくとも1つのポート34Pが機能する限り、クライアントには故障が起きていることがわからない。

【0055】HANファイルサーバ10の高可用性通信機構は、各ブレイド14ドメインに存在する通信フェイルオーバー機構(CFail)66により提供される。CFail66は、各ブレイド14のネットワーク48の機構とブレイド14A及び14Bのメッセージ42機構とについての通信故障処理のために別々に動作するものの協調的な機構を含む。

【0056】まず、ネットワーク48、すなわち、クライアント34C及び制御/プロセッササブシステム14ドメイン間の通信についてのCFail66の機能及び操作について考える。CFail66はIPバススルーと呼ばれる操作を実行し、これにより、一方のブレイド14に関連する故障したネットワーク34サービスは、反対側のピアブレイド14の故障していない対応ポート34Pに移され、以下に記述するように、ブレイド14を通る代替りのバスを介してルーティングされる。図3に示すように、各CFail66には、ブレイド14のFEP44Fに存在する通信モニタリング処理/プロトコル機構(CMonitor)66Cが含まれる。CMonitor66Cは、ブレイド14A及び14BのNetSOS56の操作と、ポート34P及びネットワーク34を介した通信と、ブレイド14A及び14B間のブレイドバス30とを介した通信を含めたブレイド14の全ての通信機能をモニタして調整する。ポート34P及びネットワーク34を介した通信のモニタリングと故障検出のために、各CFail66は、ネットワーク48とブレイド14のポート34Pとを介して動作するSLIPインターフェイス(SLIP)66Sを備えており、SLIP66Sは、ブレイド14に存在し、ネットワーク調整バケット(NCPack)66Pを反対側のピアブレイド14とやりとりする。NCPack66Pは、例えば、ネットワーク調整情報及び通知を備え、CMonitor66Cによって故障したポート34Pを検出及び識別するために用いられる。特に、各SLIP66Sは、ブレイド14間の各ネットワーク34バスを

介して、定期的に、反対側のピアブレイド14のSLIP66S及びCMonitor66CにビーコンNCPack66Pを送信する。ブレイド14のCMonitor66Cが、所定の故障検出間隔で、バスを介して反対側のピアブレイド14からビーコンNCPack66Pを受領しない場合、ブレイド14間のネットワーク34バスが、故障したものと検出される。そして、反対側のブレイド14のポート34Pインターフェイスに故障が起こったと想定される。所定故障検出間隔は、NCPack66P通信間の間隔より長く、通常CIFSクライアントタイムアウト間隔より短い。本実装においては、この間隔は、15秒のCIFSタイムアウト間隔に対し、ほぼ5秒に設定される。

【0057】図3に示すように、各CFail66は、CMonitor66Cに回答して任意のARP応答66Rを生成するARP応答ジェネレータ(ARPGen)66Gと、ネットワーク48によるクライアント34C通信のリダイレクションを管理するために、CFail66の操作にしたがってCRT48Aに存在するCRE48Eの内容を管理するバスマネージャ(PM)66Mとを含んでいる。ブレイド14のCMonitor66Cが、ポート34Pインターフェイスの故障のような、ピアブレイド14の通信バス故障を判断すると、その情報はARPGen66Gに引き渡され、ARPGen66Gは、クライアント34Cの故障箇所に割り当てられた、あるいは関連するネットワークアドレスを識別するためにARPテーブル66Tに保存された情報を使用して、故障に関係するポート34Pから接続されたクライアントへの、任意の対応ARP応答66Rを生成する。ARP応答66Rは、目標となるクライアント34CのARPテーブルの情報の修正または書き換えを行い、クライアント34Cを対応するポート34Pの動作しているポート34P、すなわち、ARP応答66Rを生成しているCFail66のポート34Pにリダイレクトする。より具体的には、ARPGen66Gにより送信された任意のARP応答66Rは、各クライアント34Cに存在するARPテーブルの修正または書き換えを行い、クライアント34Cからの通信を、ARP応答66Rを送信するARPGen66Gを含むブレイド14の対応するポート34Pに向けようとする。各CFail66は、それにより、故障した通信バスのクライアント34CをCFail66が存在するブレイド14の対応するポート34Pにリダイレクトしようとし、その結果、以下に記述するように、故障したポート34Pと通信するクライアントを機能しているポート34Pを備えたブレイド14の機能している対応ポート34Pにリダイレクトする。

【0058】さらに、各ブレイド14のPM66Mは、CMonitor66Cの操作と、ARPGen66Gによる1つ以上のARP応答66Rの生成とに、AR

P応答66Rの目標であるクライアント34Cに対応するCRT48AのCRE48Eを修正することにより応じる。特に、PM66Mは、故障したエントリ(FE)48FをARP応答が向けられていた各クライアント34Cに対応するCRE48Eに書き込んで、対応するクライアント48Cの通信がリダイレクトされたことを示し、CRT48Aにバススルーフィールド(PF)48Pを設定して、ブレイド14が1つのモードで動作していることを各ネットワーク48に知らせる。

10 【0059】この後、それ自身のポート34Pを介して、ピアブレイド14、すなわち、ピアブレイド14上でサポートされるクライアントファイルシステムあてのクライアント34Cからの通信が受領されると、ネットワーク48はPF48Pをチェックしてバススルーモード操作が有効であるかどうか判断する。バススルーモードが有効である場合、ネットワーク48は、ブレイド14のBEP44間のブレイドバス30バスからなるバススルーバスを介してピアブレイド14に通信を向ける。さらに、先に記述したリダイレクションの結果として、

20 ネットワーク48は、ブレイド14のポート34Pあてのブレイドバス30バススルーバスを介した通信ではあっても、他方のブレイド14を通るリダイレクションによりブレイドバス30バススルーバスを介してリダイレクトされた通信を受領できる。このような場合、CMonitor66C及びPM66Mは、通信ソースであつたクライアント34Cに対応するCRE48Eを修正することで、ネットワーク48による通信の受領に応じ、ブレイドバス30バススルーバス及びピアブレイド14を介してクライアント34Cに通信をルーティングする。これにより、影響を受けたクライアント34Cへの及びそれからのバスの両方向において通信のリダイレクションが完了する。

30 【0060】HANファイルサーバ10の別の実装において、各クライアント34Cは、複数のネットワーク34を介してHANファイルサーバ10のポート34Pに接続されることができ、ネットワーク34は、ローカルエリアネットワーク(LAN)、広域ネットワーク(WAN)、直接プロセッサ接続またはバス、ファイバオプティックリンク、あるいは前記の組み合わせなどの異なる技術を使用することができることを上述した。これらの実装において、CFail66機構は、ネットワーク34通信の故障が検出されると上述のように動作するが、さらに、生き残ったブレイド14にクライアント34C通信をリダイレクトすると同様に、クライアント34Cとポート34Pが故障したブレイド14との間の利用可能及び機能している代替りのネットワーク34バスを選択してもよい。この実装において、CFail66機構は、上述のように、クライアント34C ARP

40 テーブル及びCRE48Eを修正してクライアント34C通信をリダイレクトするが、代替りのバスを選択する

50

際に付加的なオプションを選択する。

【0061】上述のIPバススルー操作に関して、HANファイルサーバ10のCFail66機構が、ネットワーク34とブレイド14との間の接続場所または原因を識別しようとしないうことに注目すべきである。その代わりに、各CFail66は、反対側のブレイド14のポート34Pインターフェイスに故障が起きたと想定し、IPバススルー操作を開始する。その結果、所定の通信バスのためのIPバススルー操作が、ブレイド14A及び14Bによって同時に実行される。しかしながら、ブレイド14A及び14Bによって同時に実行されるIPバススルー操作は、本発明においては衝突しない。すなわち、例えば、バススルー操作が、ブレイド14A及び14Bの一方のポート34Pインターフェイスの故障、あるいはブレイド14A及び14Bの一方へのネットワーク34リンクの故障の結果である場合、故障に関連するブレイド14のCFail66は、そのポート34Pあるいはネットワーク34リンクを介して接続されるクライアント34CにARP応答66Rを伝達することができない。その結果、故障に関連するブレイド14のCFail66は、そのブレイド14に対応するクライアント34Cトラフィックをリダイレクトすることができない。しかしながら、反対側のブレイド14、すなわち、故障に関連しないブレイド14のCFail66は、故障したバスに関連したクライアント34CにARP応答66Rを送信し、その結果、ブレイド14に対応するクライアント34Cトラフィックをリダイレクトことに成功する。ネットワークのパーティションから生じる故障の際には、以下に記述するように、両方のポート34Pインターフェイスがブレイド14A及び14B間のブレイドバス30通信バスを介してネットワークパーティションを「橋渡し」できる。その結果、全てのクライアント34Cがブレイド14A及び14Bのどちらかと通信できる。

【0062】最後に、ブレイド14A及び14Bのどちらかが完全に故障した際には、他方のブレイド14の生き残った対応ポート34Pにより、故障したポート34Pのサービスの引き継ぎに関して上述した方法で、CFail66を介してIPバススルー操作が実行される。ただし、故障したブレイド14のポート34P全てのネットワークサービスは、生き残ったブレイド14の対応ポート34Pによって引き継がれる。しかしながら、一方のブレイド14が完全に故障してしまうと、故障したブレイド14により提供されていたクライアントのTCP接続が断ち切られてしまうので、IPバススルーの完了後再構築されなければならないことが当業者には明らかであるだろう。その後、故障したブレイド14上で利用可能だったサービスが生き残ったブレイド14上で利用可能になり、故障したブレイド14のクライアントは生き残ったブレイド14に対してTCP接続を再構築で

きる。

【0063】最後に、上述したIPバススルー機構の操作に関して、HANファイルサーバ10によってサポートされるネットワーク34関連通信操作には、上述したポイントツーポイント、またはクライアント34CからHANファイルサーバ10への通信と同様に、例えば、ネットワーク48のNetBIOS機構により、必要に応じてブロードキャスト通信が含まれることが理解されることと思う。当業者には明らかであるように、ブロードキャスト通信は、特定の受け手へというより複数の受け手にあてられる点でポイントツーポイント通信とは異なるが、ブレイド14がバススルーモードで動作している時には、クライアント34C通信に似た方法で管理される。この場合、ブロードキャスト通信を受けるネットワーク48は、上述のように、ブレイドがバススルーモードで動作しているかどうかを調べ、もしそうであるならば、ブレイドバス30バススルーバスを介して反対側のブレイド14のネットワーク48に各ブロードキャスト通信を転送する。その結果、その通信は、他のネットワーク48により直接受けたブロードキャスト通信と同様に扱われる。

【0064】上記に関して、業界標準CIFS仕様書にはクライアントシステム上で動作しているアプリケーションが接続を失った場合の影響が記載、あるいは特定されていないことが当業者にはよく知られている。経験及び実験及びアプリケーション説明書によれば、アプリケーションのTCP接続が失われた場合の影響はアプリケーションに依存しており、それぞれが故障に対して異なる処理を行う。例えば、あるアプリケーションは、クライアントにTCP接続を使用する操作を再実行するように指示し、いくつかのアプリケーションは自動的に操作を再実行する。別のアプリケーションは、ユーザに故障を報告するのみである。このため、本実装のネットワークポートフェイルオーバー機構は、これらの機能を実装するための機能を組み込んでおり、それには、各ポート34Pが複数のアドレスに対応することを可能にする、複数のIPアドレスをサポートするためにポート34Pを制御するNetDD60の機能と、故障したブレイド14からのIPアドレスを転送し、生き残ったブレイド14上のIPアドレスを作成するために必要な機能とが含まれる。ネットワークポートフェイルオーバー機構にはまた、任意のARP応答66Rを生成して故障したポート34Pに接続されたクライアントに送信し、さらにクライアントのARPテーブルのIPアドレスが新しいポート34Pをポイントするように変更したり、他のサブシステムの可用性及び故障モニタリング機能と接続してブレイド14の完全な故障がいつ起きたかを知ったり、故障したブレイド14リソース名のためのNetBIOS名の変更を行ったりする上述した機能が含まれる。

【0065】よって、HANファイルサーバ10のCFail66機構が、ブレイド14A及び14Bのポート34Pインターフェイス内のサブネットワークレベルをも含めたどのネットワークレベルに故障が起きても、クライアント34CとHANファイルサーバ10のブレイド14との間の通信を維持あるいは回復できることは明らかである。唯一の必要条件は、ブレイド14Aあるいは14Bの少なくとも一方で、1つのネットワーク通信バス及びネットワークインターフェイスが各ネットワーク34のために機能することである。従って、本発明のCFail66機構は、従来技術に典型的な、ネットワーク通信故障のソースと原因とを識別し隔離するのに必要とされる複雑な機構や手順を必要とせず、その一方でまた、衝突する可能性のある故障管理操作を調節し、同期させ、管理するのに必要とされる、これもまた従来技術に典型的な複雑な機構や操作を必要としない。

【0066】b. ブレイド14/ブレイド14通信及び故障処理機構

HANファイルサーバ10の通信機構の中間層が、ブレイドバス30及びメッセージ42のような、制御/プロセッササブシステム14ドメインのブレイド14A及び14Bドメイン間及びその内部の通信をサポートする通信機構を含むことを上述した。例えば、前述のように、ブレイドバス30バス及びメッセージ42は、ブレイド14間の一連のHANファイルサーバ10管理運営通信のために、通信引き継ぎ操作の際のファイルトランザクション操作バスのセグメントとして、CMirror54M及びLMirror54L操作においても使用される。

【0067】上述し及び図2に示すように、ブレイド14間のブレイドバス30通信バスは、ブレイドバス30、及び、各ブレイド14のBEP44Bに存在するBE Bus Sys 380から構成され、BE Bus Sys 380には、Ser-Des 38R、FCC 38Q、P-Pブリッジ38H、MCH 38C、プロセッサ36Aなどのエレメントが含まれる。図2には示されていないものの、BE Bus Sys 380はまた、プロセッサ36Aで、すなわち、BEP 44Bで動作するBE Bus Sys 380制御通信機構を備えている。BE Bus Sys 380制御通信機構は、通常、当業者には明らかな方法で動作し、BE Bus Sys 380及びブレイドバス30を介する通信操作を実行する。プロセッサ36A及び36B、すなわち、各ブレイド14のFEP 44F及びBEP 44Bはまた、図2あるいは3に示されていないメッセージ42制御通信機構を実行することが理解されるだろう。メッセージ42制御通信機構は、通常、当業者には明らかな方法で動作し、メッセージ42を介する通信操作を実行する。

【0068】BEP 44B及びFEP 44A間の通信を提供するメッセージ42は、各ブレイド14のメモリ3

8Aの共有メッセージ通信空間と、プロセッサ36A及び36Bで動作するメッセージング機構とから構成される。メッセージング機構は、通常、当業者には明らかな方法で動作し、メッセージ42を介する通信操作を実行する。

【0069】図3に示すように、CFail66には、SLIP66S、CMonitor66C、ARPGen66Gとは別の独立した故障処理機構が含まれる。SLIP66S、CMonitor66C、ARPGen66Gは、制御/プロセッササブシステム14ドメインのブレイド14A及び14Bドメイン間及びその内部の通信についての故障処理のために、制御/プロセッササブシステム14ドメインへの及びそれからの通信と関連して機能する。図からわかるように、CFail66の相互ブレイド14ドメイン通信故障処理機構には、ブレイドバス30及びブレイド14のBE Bus Sys 380を含めた、ブレイド14A及び14B間のブレイドバス30通信リンクの操作をモニタするブレイド通信モニタ(BMonitor)66Bと、ブレイド14のメッセージ42の操作とが含まれる。しかしながら、この接続は図3には示されていない。まずブレイドバス30を取り上げると、ブレイド14間、すなわち、ブレイドバス30あるいはBE Bus Sys 380のブレイドバス30通信バスが何らかの理由で故障すると、この故障はBMonitor66Bによって検出され、通常、プロセッサ36Aで動作するBE Bus Sys 380制御機構が、ブレイドバス30バスを介して試みられた通信が受領確認されていないと通知する。

【0070】ブレイドバス30通信バスの故障の際には、BMonitor66Bは、ブレイド14A及び14B間の利用可能な通信ルーティングバスに関する情報を保存しているブレイドルーティングテーブル(BRT)48Bを読み取る。そこに保存されたバス情報は、例えば、ブレイドバス30を介する通信のルーティング情報を含み、さらに、ブレイド14A及び14B間の利用可能なネットワーク34バスのルーティング情報も含む。BRT48BはCFail66に関連して保存されるが、図3に示すように、本実施例のブレイド14においては、BRT48Bはネットワーク48と関連して存在する。そのため、ネットワーク34に関連するルーティングバス情報はすぐに利用されることができ、CRT48Aの構築などのネットワーク48の正常動作時にはネットワーク48にアクセスすることができる。BMONITOR66Bは、故障したブレイドバス30のバスを除いて、ブレイド14間の利用可能な通信バスについてのルーティング情報を読み取り、ブレイドバス30バスの後継あるいは代理で使用される、ブレイド14のネットワーク48間の利用可能なネットワーク34バスを選択する。この関係で、BMONITOR66Bが、PM66MがCRT48AのCRE48Eを修正するのと

同様かつ同時に、全てのIPバススルー操作の間にBRT48Bの内容を修正して、ブレイド14間の機能していないネットワーク34バスを示すことに注意しなければならない。この結果、ブレイドバス30バスの後継バスは、機能しているネットワーク34バスのみから選択される。

【0071】BMonitor66Bは、その後、FEP44F及びBEP44Bで動作するBE Bus Sys 380及びメッセージ42制御通信機構に、ブレイドバス30バスにルーティングされる全ての通信を、BEP44Bにより直接、あるいはFEP44Fによりメッセージ42を介して間接的に、ネットワーク48及びPM66Mにより選択されたネットワーク34バスへリダイレクトするという通知を出す。

【0072】従って、どんな理由によりブレイド14間のブレイドバス30通信バスに故障が起きても、CFail66のCMonitor66C及びBMonitor66B機構は、ネットワーク34を介してブレイド14からブレイド14への通信のために代わりの通信バスを見つけて使用できる。この関係で、CFail66機構が、故障の場所あるいは原因を識別しようとしないうで、故障のソースを識別して隔離するのに通常必要となる複雑な機構及び手続と、衝突する可能性のある故障管理操作を調整し、同期させ、管理するのに通常必要となる複雑な機構及び操作とを必要としないことに再び注目すべきである。

【0073】また、HANファイルサーバ10の通信故障処理機構は、互いに別個に独立して動作するが、これによりまた、衝突する可能性のある故障管理操作を調節し、同期させ、管理するための複雑な機構及び操作を利用する必要がなく、複数の故障ソースあるいは複数の故障を協調して処理できることに注目しなければならない。例えば、CFail66ネットワーク34故障機構、すなわち、CMonitor66C関連機構によって実行される操作は、CFail66ブレイドバス30故障機構、すなわち、BMonitor66B関連機構によって実行される操作とは別に実行されるが、クライアント34C及びブレイド14間、そしてブレイド14間の通信を維持するために機能的に協調して実行される。ブレイド14間の、そして各クライアント34Cへのネットワーク34バスが、ブレイドバス30バスが故障を起こした時に、1つでも機能していれば、通信は、故障のソースあるいは故障の順番に関わらず維持される。

【0074】例を示すと、第一ブレイド14と関連するネットワーク34に故障が起きると、上述のように、第二ブレイド14を介しての、そしてCFail66ネットワーク34故障機構によりブレイド14間のブレイドバス30リンクを介しての第一ブレイド14への、クライアント34C通信のリダイレクションが生じる。次に

ブレイドバス30リンクに故障が起これと、CFail66ブレイドバス30故障機構により、第二及び第一ブレイド14間で機能している代わりのネットワーク34バスを介して、第二ブレイド14及びブレイドバス30リンクを介してリダイレクトされたクライアント34通信が再び、第二ブレイド14から第一ブレイド14へリダイレクトされる。

【0075】さらなる例では、第一の故障がブレイドバス30リンクで起きた場合、ブレイド14間の通信は、上述のように、CFail66ブレイドバス30故障機構により、ネットワーク34を介してブレイド14間で機能している代わりのバスへリダイレクトされる。この代わりのネットワーク34バスにおいて次なる故障が起きた場合、この故障はネットワーク34関連の故障として検出され、ブレイド14のCFail66ネットワーク34故障機構は、まず、ブレイドバス30リンクを介してブレイド14間の先にリダイレクトされた通信をルーティングしようとする。しかしながら、CFail66ブレイドバス30故障機構は、ブレイドバス30リンクが機能していないために、ブレイド14間の利用可能で機能している代わりのネットワーク34バスを介して先にリダイレクトされた通信をリダイレクトする。

【0076】従って、ネットワーク34及びブレイドバス30の故障がどんな組み合わせあるいは順番で起こっても、クライアント34Cとブレイド14との間、そしてブレイド14間の通信を維持するために、CFail66ネットワーク34及びブレイドバス30故障機構がさまざまな組み合わせ及び順番で別個の独立した操作を実行することが明らかであろう。また、ブレイドバス30バスに故障が起きた際に、ブレイド14間、そして各クライアントへのネットワーク34バスがたった1つでも機能している限り、故障のソースあるいは故障の順番に関係なく通信は維持される。

【0077】最後に、この関係で、ブレイド14のFEP44F及びBEP44B間のメッセージ42リンクに故障が起きる可能性があることに注意しなければならない。多くの場合、これはブレイド14が完全に故障した結果であるが、幾つかの場合において、故障はメッセージ42機構に限定されることが出来る。メッセージ42機構に限定された故障の場合、故障が起きたブレイド14のFEP44Fは、ブレイド14のBEP44Bと、あるいは反対側のブレイド14と通信することができなくなり、BEP44BはブレイドのFEP44Bと通信できなくなるが、ブレイド14間のブレイドバス30リンクを介して反対側のブレイド14のBEP44B及びFEP44Fと通信できる。

【0078】従って、本発明のさらなる実装においては、メッセージ42に故障が起きたブレイド14のBMonitor66Bは、FEP44Fに関連してブレイドバス30の明らかな故障を検出するが、BEP44B

に関連するブレイドバス30の故障を検出しない。従って、このブレイド14のBMonitor66B及びCMonitor66C機構は、PM66Mによって選択されたネットワーク34バスを介して、FEP44Pから全ての通信をBEP44Bへ、あるいは反対側のブレイド14ヘリダイレクトし、BEP44BからFEP44Fへの全ての通信をブレイドバス30、及びFEP44Fのために選択されたネットワーク34バスを介するルートヘリダイレクトするが、ブレイドバス30を介するBEP44B通信をリダイレクトしない。

【0079】故障が起きなかったブレイド14においては、BMonitor66B機構は、メッセージ42が故障したブレイド14のFEP44Pへの通信について明らかなブレイドバス30バス故障を検出するが、そのブレイド14のBEP44Bへの通信についてのブレイドバス30バス故障を検出しない。従って、このブレイド14のBMonitor66B及びCMonitor66C機構は、反対側のブレイド14のFEP44Fあての全ての通信を、上述のように、代替りのネットワーク34バスを介してリダイレクトするが、反対側のブレイド14のBEP44Bあての通信をリダイレクトしない。

【0080】c. 記憶サブシステム12/ブレイド14故障処理機構

上述のように、HANファイルサーバ10の故障処理機構の最下層レベルには、記憶サブシステム12の通信バス構造及びRAID46によって提供されるRAIDF46F機構とが含まれる。RAIDファイル機能は、当業者にはよく知られているため、ここでは本発明を理解するのに必要な場合のみ説明し、以下には、記憶サブシステム12内部の、そしてサブシステム12及びブレイド14間の通信バスに焦点を当てて説明する。

【0081】図1に示すように、そして上述したように、記憶サブシステム12には複数のハードディスクドライブ18から構成されるドライブバンク16が含まれる。各ハードディスクドライブ18は、デュアル記憶ループモジュール20A及び20Bを介して双方向に読み取り/書き込みアクセスされる。記憶ループモジュール20A及び20Bそれぞれには、MUXBANK22A及び22Bが含まれ、各MUXBANK22には、複数のMUX24とループコントローラ26A及び26Bとが含まれる。各ループコントローラモジュール20のMUX24とループコントローラ26とは、MUXループバス28A及び28Bを介して双方向に相互接続されている。図からわかるように、MUXBANK22A及び22Bそれぞれには、対応するディスクドライブ18の1つに対応して接続されるMUX24Dが含まれる。そのため、ドライブバンク16の各ディスクドライブ18は、MUXBANK22A及び22Bそれぞれの対応するMUX24Dに接続されて双方向に読み取り/書き込

みされる。MUXBANK22A及び22Bそれぞれには、さらに、MUX24CA及びMUX24CBを介して対応するコンピュータブレイド14A及び14Bの一方が双方向に接続されており、コンピュータブレイド14A及び14Bは、ブレイドバス30を介して双方向に接続されている。

【0082】従って、各ディスクドライブ18は、MUXバンク22AのMUX24DとMUXバンク22BのMUX24Dとに双方向に接続されている。MUXバンク22AのMUX24は、ループバス26Aを介して相互接続されている一方、MUXバンク22BのMUX24は、ループバス26Bを介して接続されている。そのため、各ディスクドライブ18は、ループバス26A及びループバス26B両方を介してアクセス可能である。さらに、プロセッサブレイド14Aは、ループバス26Aと双方向に通信する一方、プロセッサブレイド14Bは、ループバス26Bと双方向に通信し、プロセッサブレイド14A及び14Bは、ブレイドループ（ブレイド）バス30を介して直接相互接続されて通信する。

【0083】従って、記憶サブシステム12内部の下層レベルの通信故障処理機構が、基本的に、各ディスクドライブ18とプロセッサブレイド14A及び14Bとの間に複数の予備のアクセスバスを提供する受動的なバス構造であることがわかるだろう。このため、プロセッサブレイド14A及び14Bは、記憶サブシステム12内部の1つ以上の通信バスで故障が起きた際には、対応するループバス26を介して直接、あるいは他方のプロセッサブレイド14を介して間接的に、ディスクドライブ18のどれとでも双方向通信が可能であり、互いに直接通信できる。1つ以上のディスクドライブ18内で起きる故障のための故障処理機構は、上述のRAIDF48F機構から構成される。

【0084】また、記憶サブシステム12の受動バス構造が、通信機構と、ブレイド14のCFail66ネットワーク34及びブレイドバス30故障機構とは別々に独立して動作するものの、クライアント34Cと、クライアント34のファイルシステムが存在するディスクドライブ18との間の通信を保証するために、ブレイド14の機構と協調して動作することがわかるだろう。また、これらの機構は、複雑な故障検出、識別、隔離機構の利用と、複雑な故障管理調整、同期、管理機構の利用とを廃して、高レベルのファイルシステム可用性を提供する。

【0085】5. HANファイルサーバ10のファイルトランザクション故障処理機構とHANファイルサーバ10の通信故障処理機構の相互運用（図1、2、3）
本実施例のHANファイルサーバ10が、多数の高可用性機構、すなわち、HANファイルサーバ10の1つ以上のコンポーネントに故障が起きた際にも、HANファイルサーバ10がクライアントへのファイルサーバサ

ビスを中断せずに提供し続けることを可能にする機構を備えることを上述した。これらの機構の多くは、基本RAID F46 F機能のように、従来技術の代表的なものであり、当業者にとっては周知のものである。そのため、本発明に関係しない限り詳細な説明を省く。

【0086】しかしながら、一般的には、HANファイルサーバ10のコンポーネントに故障が起きた際には、HANファイルサーバ10の生き残ったコンポーネントが、高可用性機構の操作により、故障したコンポーネントによって実行されていたタスク及びサービスを引き継ぎ、これらのサービスの提供を続ける。このような高可用性機構の操作には数多くの機能があり、そのような機構がこれらの機能を達成するためには幾つかの操作を実行する必要があることが当業者には明らかであろう。例えば、高可用性機構は、コンポーネントの故障を識別し、故障したコンポーネントから生き残ったコンポーネントへソースあるいは機能の引き渡しあるいは移転を行い、故障したコンポーネントによって提供されていたサービス及び機能が外からわかるように中断されないように生き残ったコンポーネントに引き継がれたリソースの状態を回復し、故障したコンポーネントの置換あるいは訂正を行ない、修復後には故障していたコンポーネントにリソースを引き渡すあるいは移動する必要がある。

【0087】通信に関して上述したように、HANファイルサーバ10のファイルトランザクション及び通信機構は、独立して動作する。そして以下にさらに詳細に説明されるように、本発明のHANファイルサーバ10の高可用性機構は、HANファイルサーバ10の多数の異なる機能レベルで動作する。通常、異なるグループ、あるいは異なるタイプの操作及び機能は、HANファイルサーバ10の各機能レベルで実行される。従って、高可用性機構はそれぞれ異なり、各レベルで、そしてシステムとしてのHANファイルサーバ10のために、独立しながらも協調して動作して高レベルのサーバ可用性を提供する。以下にさらに詳細にこれらの機構の構造及び操作と、これらの機構の相互運用とを説明する。

【0088】例えば、HANファイルサーバ10における最上層レベルの機能は、クライアント通信タスク及びサービスを実行する通信レベル、すなわち、クライアントと、ネットワーク34を介してHANファイルサーバ10によってサポートされるクライアントファイルシステムとの間の通信である。この通信レベルの中心機能は、ネットワーク48の機構とHANファイルサーバ10の関連コンポーネントとによって提供される。通信レベルでの高可用性機構には、CFail66のような故障検出機構が含まれ、通信レベルでの故障を処理する多数の異なる機構を提供する。例えば、ブレイド14A及び14Bのうちの一方で1つ以上のポート34Pを介する通信に故障が起きた場合、ピアブレイド14のCFail66は故障を検出し、ネットワーク48と連携し

て、クライアントと故障したポート34Pとの間の全ての通信を、ピアブレイド14の機能している対応ポート34Pにリダイレクトする。ピアブレイド14では、その内部のネットワーク48が、ブレイドバス30を介して、故障したポート34Pを有するブレイド14のJFile50に通信をルーティングする。その結果、故障したポート34Pは、ピアブレイド14のポート34Pと、ブレイドバス30及びメッセージ42を介するFEP44F-BEP44P通信バスからなる相互ブレイド14通信バスとを介してバイパスされる。この関係で、ブレイド14の高レベルファイルトランザクション機構について以下の記述により説明されるように、ネットワーク48の高可用性機構は、高レベルファイルトランザクション機構の高可用性機構を相互運用して、実際の、そして例えば、ブレイド14JFile50のあるいはブレイド14全体の故障から生じる明らかなネットワーク34関連通信故障に対処する。

【0089】ブレイド14における次のレベルの機能は、高レベルファイルトランザクション機能及びサービスから構成される。そこでは、高レベルトランザクション機能の中心機能及び操作は、JFile50及び関連する高レベルファイル機構により提供される。上述のように、HANファイルサーバ10の高レベルファイル機能レベルでの高可用性機構には、CMirror54Mを備えたWCACHE50CとLMirror54Lを備えたログ50Lとが含まれ、これらの機構は、ブレイド14内部の高レベルファイル機構の故障を処理する。上述のように、WCACHE50Cは、従来方法で動作してデータトランザクションをキャッシュし、CMirror54Mは、WCACHE50Cに影響するFEP44Fに故障が起きた際、WCACHE50Cの内容を回復できる。ログ50Lは、ブレイド14とともに動作してJFile50により実行されるファイルトランザクションの履歴を保存する。これにより、ログ50Lは、例えば、トランザクションが記憶サブシステム12の固定記憶装置に完全にコミットされる前にファイルトランザクションの損失を生じる、JFile50あるいは記憶サブシステム12の故障の際、失われたファイルトランザクションを再実行及び回復させることができる。

【0090】しかしながら、LMirror54L機構は、LMirror54Lがミラーリングするログ50Lが存在するブレイド14内部で動作せず、代わりに、ブレイド14を横断して動作して、各LMirror54Lが、反対側のピアブレイド14のログ50Lの内容をミラーリングして保存できるようにしている。その結果、LMirror54L機構は、反対側のピアブレイド14に壊滅的な故障が起きた場合にも反対側のピアブレイド14のログ50Lの内容を保存し、故障していたブレイド14がサービスを再開した際に、失われたファ

イルトランザクションを故障していたブレイド14で再実行及び回復することができる。

【0091】さらに、生き残ったブレイド14内部に故障したブレイド14の失われた可能性のあるファイルトランザクションのレジデント履歴を備えることにより、LMirror54L機構はまた、生き残ったブレイド14に故障したブレイド14によってサポートされていたクライアントのサポートを引き継がせることができることに注目すべきである。すなわち、ネットワーク48機構について上述したように、生き残ったブレイド14のネットワーク48及びJFile50は、故障したブレイド14のクライアントを生き残ったブレイド14にリダイレクトすることにより、故障したブレイド14によって先にサポートされていたクライアントのサービスを引き継ぐ。この処理では、上述のように、生き残ったブレイド14のネットワーク48機構は、生き残ったブレイド14のJFile50に、引き継がれたIPアドレスあてのデータトランザクションを向けることにより、故障したブレイド14のIPアドレスを引き継ぐ。生き残ったブレイド14のJFile50は、生き残ったブレイド14がローカルファイルシステムを備えるという仮定の下に、新しいクライアントとして故障したブレイド14のクライアントを引き継ぎ、その後は、引き継がれたクライアントを自分のクライアントとしてサービスを行う。そのサービスには、引き継がれたデータトランザクションを処理することと並行して全ての引き継がれたデータトランザクションを記録することが含まれる。生き残ったブレイド14は、ローカルリカバリログ、すなわち、生き残ったブレイド14に存在するLMirror54Lを使って引き継いだIPアドレスのデータトランザクションを記録するとともに、レジデントLMirror54Lに保存されたファイルトランザクション履歴を使用して故障したブレイド14の失われたファイルトランザクションを再実行及び再構成し、故障したブレイド14のクライアントのファイルシステムを所望の状態に回復することができる。この関係で、生き残ったブレイド14のJFile50は、故障したブレイド14に向けられていたファイルトランザクションの初期アドレスを基にしてネットワーク48からの通知により、あるいはレジデントLMirror54Lの内容を調べて保存されたファイルトランザクションと相互に関連する「新しい」クライアントファイルトランザクションがあるかどうか判断することにより、「新しい」クライアントが故障したブレイド14から移転されたクライアントであるかを判断できる。

【0092】最後に、HANファイルサーバ10の最下層レベルのファイルトランザクション機能は、RAID46によってサポートされるRAID46ファイルトランザクション機能及びサービスから構成される。RAID46F機能は、それ自身、上層レベルの高可用性機

構から独立して動作することがわかるだろう。しかしながら、通信レベル及び高レベルファイルトランザクション機構は、例えば、デュアルブレイド14A及び14B、ループバス26A及び26B、MUXループバス28A及び28Bを介する代わりの通信バスの提供と連携してRAIDF46F機能と協調的に動作し、ディスクドライブ18へのアクセス可能性を高めていることがわかるだろう。

【0093】従って、HANファイルサーバ10に設けられた通信レベル及び高レベルファイルトランザクション機構と代わりの通信バスとは、RAIDF46F機能と協力してネットワーククライアントへのファイルシステム共有資源、すなわち、記憶空間の可用性を高めることが上記より理解されることができ。また、HANファイルサーバ10に設けられた通信レベル及び高レベルファイルトランザクション機構と代わりの通信バスとが、複雑な故障検出、識別、隔離機構の利用、及び複雑な故障管理調整、同期、管理機構の利用を廃して、上記の効果を達成することが理解されるだろう。

【0094】よって、要約すると、数多くの異なる機構が故障したコンポーネントを識別するために用いられ、その機構は、コンポーネントと、コンポーネントが存在するHANファイルサーバ10のサブシステムと、コンポーネントの故障によるHANファイルサーバ10の操作への影響とに依存して特定されることが上記から理解される。例えば、RAIDM46M機能が、ファンや電源装置のようなコンポーネント、及びブレイド14A及び14Bの類似のコンポーネントの故障をモニタして検出する一方、RAIDF46F機能は、ディスクドライブ18のファイルシステム操作のエラー及び故障をモニタ、検出、修正あるいは補正する。RAID46機構によってモニタされるコンポーネントの多くは故障が起きても、システムとしてのHANファイルサーバ10レベルでのデータの可用性を危うくすることはないが、そのコンポーネントを修復するための処置を取ることができるように管理インターフェースを通じて検出及び連絡されなければならないことがわかるだろう。さらなる例では、HANファイルサーバ10のネットワーク管理機能は、ネットワーク34の状態と、HANファイルサーバ10のネットワーク34通信関連コンポーネントとをモニタし、それぞれの故障に適した方法で、HANファイルサーバ10とHANファイルサーバ10のクライアントとの間での通信の故障に対応する。ネットワークをモニタするために、ネットワーク管理機能は、HANファイルサーバ10自身のネットワーク通信をテストするためのセルフチェックを生成し、外部ネットワークと通信しているかどうか判断する。例えば、このセルフチェックがネットワークバスのどれかで失敗する場合、故障したネットワークバスによってサポートされていた通信は、上述のように別のネットワークバスに引き継がれ

る。さらに別の例においては、RAID46機能がブレイド14の故障を検出すると、この故障が上述のようにファイルシステム機能に連絡され、その結果、フェイルオーバー処理が適切なファイルシステムレベルで実行されることができる。

【0095】故障処理過程での次のステップ、すなわち、生き残ったリソースへの故障したリソースの移転は、通常、既知の生き残った場所にリソースを再割り当てすることにより実行される。ネットワーク機能の故障の場合、移転は、上述のように、故障したデバイスの機能を引き継ぐことのできる、先に識別されたネットワークアダプタに対して行われる。故障したのがブレイド14である場合は、ピアブレイド14が故障したブレイド14からファイルシステムを引き継ぐ。

【0096】故障したコンポーネントから生き残ったコンポーネントへのリソースの移転には、そのリソースが生き残ったコンポーネント上で利用可能にされる前にリソースの動作状態を変更あるいは修正する必要がある。例えば、ネットワークコンポーネントの故障の場合、新しいネットワークアドレスが既存のアダプタに付加されなければならない、ブレイド14の故障のようにファイルシステムに影響を与える故障の場合には、トランザクションログを再実行して故障で失われたデータを置換する。

【0097】先に記述したように、HANファイルサーバ10のコンポーネントの多くは、HANファイルサーバ10から取り外して、動作しているコンポーネントに置換することができる、ホットスワップ可能なコンポーネントである。一旦コンポーネントを置換すると、生き残ったコンポーネントにより引き継がれたリソースは初期のコンポーネントに、つまりは、初期のコンポーネントが置換されたものに戻されなくてはならない。従って、上述のような適切なサブシステムの回復機構では、生き残ったコンポーネントに移転されたリソースは置換されたコンポーネントに移行される。この処置は、通常、システムアドミニストレータにより手動で、そしてサービスの中断が受け入れ可能及び処理可能な時に行なわれる。

【0098】本発明が、ここに例として使われたファイルサーバと同様に、例えば通信サーバ、さまざまなタイプのデータプロセッササーバ、プリンタサーバなどの、クライアントとの信頼できる通信と、データあるいは処理トランザクションの保存及び回復とを必要とするあらゆる形式の共有リソースに実装可能であることが当業者には明らかであろう。また、本発明が、例えば、異なるRAID技術、異なる保存技術、異なる通信技術、そして画像処理などの他の情報処理手法及び技術を使用するファイルサーバの実装にも、同様に適応できるとともに実装可能であることが明らかであろう。異なる形式の共有リソース、異なるリソースマネージャ、異なるシステ

ム構成及びアーキテクチャ、異なるプロトコルにも本発明が適応できることは当業者には明らかであろう。

【0099】従って、本発明が、実施例の装置及び方法について特に説明され記述されてはいても、ここに説明され、付属の請求項によって規定される本発明の範囲を超えない限り、形式、詳細、実装におけるさまざまな変更、変形、修正を本発明に加えることができることが当業者には明らかであろう。よって、本発明のあらゆる変形及び修正を本発明の範囲内に収まるようにカバーすることが付属の請求項の目的である。

【図面の簡単な説明】

【図1】 本発明が実装されることのできるネットワークファイルサーバのブロック図である。

【図2】 図1のファイルサーバのドメインにおけるプロセッサのコアのブロック図である。

【図3】 図1のファイルサーバのドメインをさらに詳細に示した概略図である。

【符号の説明】

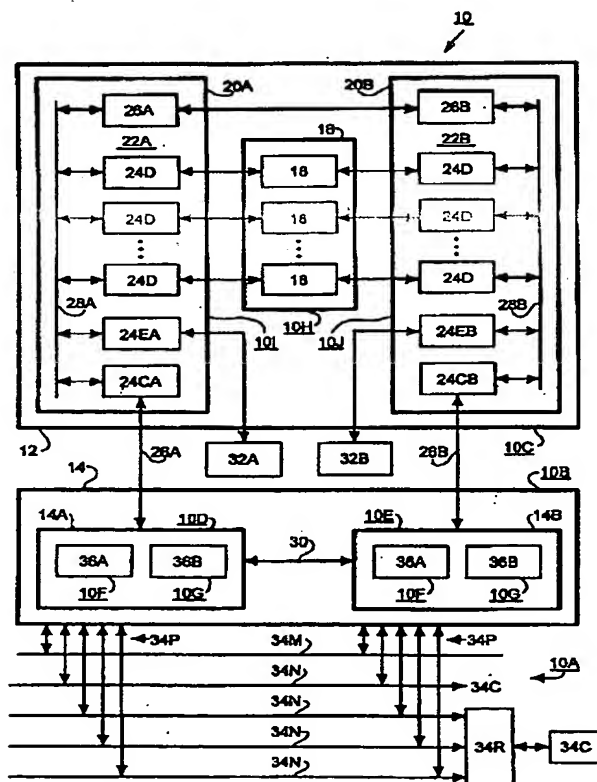
10	HANファイルサーバ
12	記憶サブシステム
14	制御/プロセッササブシステム
14A、14B	プロセッサブレイド
16	ドライブバンク
18	ディスクドライブ
20A、20B	記憶ループモジュール
22A、22B	マルチプレクサバンク
26A、26B	ループコントローラ
28A、28B	MUXループバス
30	ブレイドバス
32A、32B	外部ディスクアレイ
34C	クライアント
34M	管理ネットワーク
34N	クライアントネットワーク
34P	ネットワークポート
34R	ルータ
36A、36B	処理ユニット
38C	メモリコントローラハブ
38D	メモリ
38E	入出力コントローラハブ
38F	フロントサイドバス
38G	ハブリンクバス
38H	P-Pブリッジ
38I	ファームウェアメモリ
38J	ハードウェアモニタ
38K	ブートドライブ
38L	スーパーI/Oデバイス
38M	VGAデバイス
38N	ネットワークデバイス
38O	バックエンドバスサブシステム
38P	フロントエンドバスサブシステム

可用性が高い通信を提供する通信バススルー共有システムリソース、ネットワークファイ...

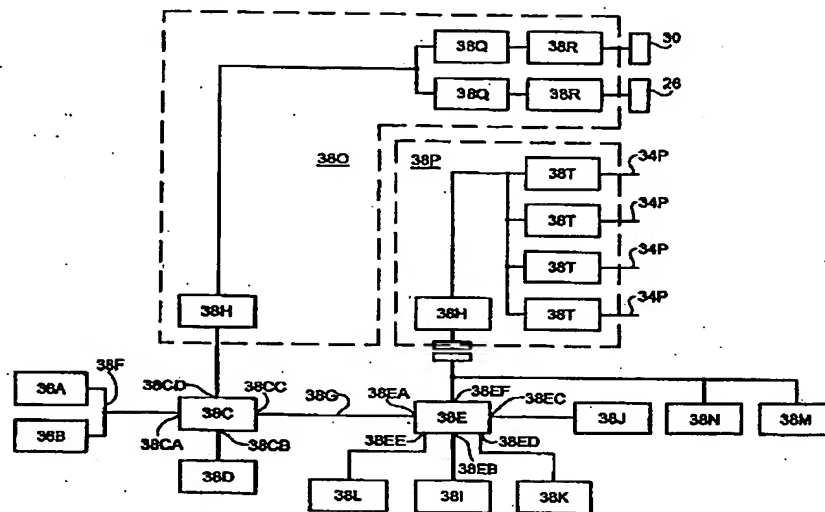
特開2002-41348

38Q	ファイバチャネルコントローラ	50L	トランザクションログ
38R	シリアルライザ/デシリアルライザデバイ	50M	ログメモリ
ス		54L	ログミラー機構
38T	ネットワークデバイス	54M	キャッシュミラー機構
40	オペレーティングシステム	05 56	ネットワークスタックオペレーション
42	メッセージパッシング機構	グシステム	
44B	バックエンドプロセッサ	58	TCP/IPプロトコルスタック
44F	フロントエンドプロセッサ	60	ネットワークデバイスドライバ
46	RAID機構	62	CIFS
46M	RAIDモニタ機構	10 64	NFS
46F	RAIDファイル機構	66	通信フェイルオーバー機構
48	ネットワーク機構	66B	ブレイド通信モニタ
48A	クライアントルーティングテーブル	66C	通信モニタリング処理/プロトコル機
48B	ブレイドルーティングテーブル	構	
48E	クライアントルーティングエントリ	15 66G	ARP応答ジェネレータ
48P	バススルーフィールド	66M	バスマネージャ
50	ジャーナルファイルシステム	66P	ネットワーク調整パケット
50C	書き込みキャッシュ	66R	ARP応答
50F	ファイルシステム機構	66S	SLIPインターフェイス
50G	ログジェネレータ	20	

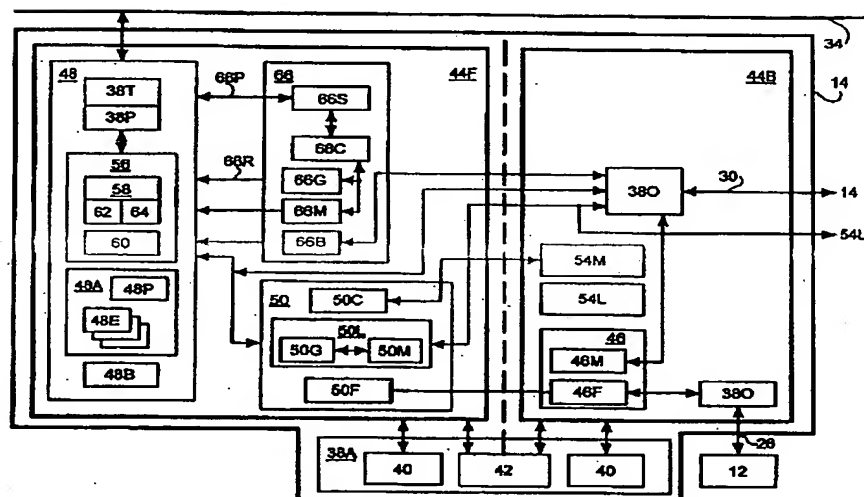
【図1】



【図 2】



【図 3】



フロントページの続き

(72) 発明者 ジェームズ グレゴリー ジョーンズ
アメリカ合衆国 ノースカロライナ州
27615 ローリー モントーク ドライブ
8708

Fターム(参考) 5B082 DD00 DE02
5B083 AA08 BB01 CD11 EE11
5B089 GA12 JB17 KA12 KB02 KC15
KG05 KG08 ME02 ME04

(19)



JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: 2000242434 A

(43) Date of publication of application: 08.09.00

(51) Int. Cl.

G06F 3/06

(21) Application number: 11344260

(22) Date of filing: 03.12.99

(30) Priority: 22.12.98 JP 10364079

(71) Applicant: HITACHI LTD

(72) Inventor:
 MATSUNAMI NAOTO
 OEDA TAKASHI
 YAMAMOTO AKIRA
 AJIMATSU YASUYUKI
 SATO MASAHIKO

(54) STORAGE DEVICE SYSTEM

COPYRIGHT: (C)2000,JPO

(57) Abstract:

PROBLEM TO BE SOLVED: To construct a storage device system corresponding to the scale or request of a computer system so that the extension of a storage device system and improvement in reliability in the future are easily realized.

SOLUTION: This system 1 has a plurality of subsets 10 having a storage device for holding data and a controller for controlling the storage device and switch devices 20 arranged between the subsets 10 and a host 30. Each switch device 20 has a managing table for holding management information for managing the configuration of the storage device system 1. According to the management information, address information contained in frame information outputted by the host 30 is translated and the frame information is distributed to the subsets 10.

